Vol. 50 DECEMBER 1959 No. 6

The Journal of Educational Psychology

CONTENTS

Some Behavioral Correlates of Teacher Effectiveness 239
DONALD M. MEDLEY AND HAROLD E. MITZEL

The Development of Understanding in Arithmetic by a Teaching Machine 247
EVAN R. KEISLAR

Transfer from Context by Subthreshold Summation 254
GEORGE M. HASLERUD

Adult Age Differences in Performance on a Visual-Spatial Task of Stimulus Generalization 259
FRANKLYN N. ARNHOFF

Scholastic Behavior and Orientation to College 266
ROBERT C. BIRNEY AND MARC J. TAYLOR

The Relationship between Salary Policies and Teacher Morale 275
CLAUDE MATHIS

Dimensions of Leadership Behavior in Classroom Discussion Groups . 280
CLAUDE J. BARTLETT

Individual Differences in Memory 285
JAMES B. STROUD AND LOWELL SCHOER

A Note on the Effect of Filling Out an "Anxiety Scale" on Examination Performance 293
NEIL A. CARRIER

Psychological Health and Classroom Functioning: A Study of Dissatisfaction with School among Adolescents 295
PHILIP W. JACKSON AND JACOB W. GETZELS

Note Taking During or After the Lecture 301
SIGMUND EISNER AND KERMIT ROHDE

This is the last issue of Volume 50
Volume contents and title page appear herein.

AMERICAN PSYCHOLOGICAL ASSOCIATION

The Journal of Educational Psychology

RAYMOND G. KUHLEN, Editor
Syracuse University

Advisory Editors

LEE J. CRONBACH
University of Illinois
FRANCIS J. DIVESTA
Syracuse University
CHESTER W. HARRIS
University of Wisconsin
ERNEST R. HILGARD
Stanford University
NICHOLAS HOBBS
George Peabody College For Teachers
HAROLD E. JONES
University of California
IRVING LORGE
Teachers College, Columbia University
DOROTHEA MCCARTHY
Fordham University
T. ERNEST NEWLAND
University of Illinois
H. H. REMMERS
Purdue University
JULIAN C. STANLEY
University of Wisconsin
J. B. STROUD
State University of Iowa
DONALD E. SUPER
Teachers College, Columbia University
PERCIVAL M. SYMONDS
Teachers College, Columbia University
GEORGE G. THOMPSON
Ohio State University

The Journal of Educational Psychology publishes original investigations of problems of learning, teaching, and the psychological development of the individual. It contains articles on the following subjects: the psychology of school subjects; experimental studies of learning; the development of interests, attitudes, personality, social relations, and vocational orientation, particularly (but not exclusively) as related to school adjustment or the impact of the school program upon development and adjustment; emotion, motivation, and character; mental development; and methods. The last includes tests, statistical techniques, and research techniques in cross-sectional and developmental studies. Studies published deal with all levels of education—preschool, elementary, secondary, college, and adult education—and with all age groups.

Manuscripts and correspondence on editorial matters should be addressed to the Editor: Raymond G. Kuhlen, Department of Psychology, Syracuse University, 123 College Place, Syracuse 10, New York.

All manuscripts should be submitted *in duplicate* to facilitate referral to advisory editors and must be prepared to conform to the style and procedures described in the *Publication Manual of the American Psychological Association*. Because of the great demand for journal space, authors should give special attention to reducing the length of manuscripts while maintaining clarity of expression. Ordinarily, papers should not exceed 15 typed pages. Accepted manuscripts are published in order of their receipt. An author may, however, request early publication of accepted manuscripts if he is willing to assume the cost of printing the necessary extra pages.

ARTHUR C. HOFFMAN
Managing Editor

HELEN ORR
Promotion Manager

Communications including subscriptions, orders of back issues, and changes of address should be addressed to the American Psychological Association, 1333 Sixteenth Street, N.W., Washington 6, D. C. Address changes must reach the APA by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Annual subscription: $8.00 (Foreign $8.50). Single copies, $1.50.

PUBLISHED BIMONTHLY BY:

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Mt. Royal & Guilford Aves., Baltimore 2, Md.
and 1333 Sixteenth Street N.W., Washington 6, D. C.

Second-Class postage paid at Baltimore, Md.
Copyright, 1959 by the American Psychological Association, Inc.

THE JOURNAL OF

EDUCATIONAL PSYCHOLOGY

Volume 50 December 1959 Number 6

SOME BEHAVIORAL CORRELATES OF TEACHER EFFECTIVENESS[1]

DONALD M. MEDLEY AND HAROLD E. MITZEL

Division of Teacher Education, Municipal Colleges of New York City

The Research Office of the Division of Teacher Education of the City of New York undertook, beginning in 1953, a longitudinal study of graduates of the teacher education program in the four municipal colleges—City, Hunter, Brooklyn, and Queens. This study has followed the general plan proposed by the Committee on Criteria of Teacher Effectiveness created by the American Educational Research Association in 1950 (Barr, Bechdolt, Coxe, Gage, Orleans, Remmers, & Ryans, 1952). The Committee suggested three steps: the definition and measurement of dimensions of teacher behavior; and the study of relationships between these two sets of dimensions. The purpose of this study was to carry out a phase of the third step listed above by examining the relationships between some measures of teacher effectiveness and some teacher behavior variables obtained in the course of an investigation of the graduates of a coordinated teacher preparation program in the municipal colleges of New York City.

Sample

In the fall of 1954, a follow-up was made to locate those graduates from the student teaching class of 1953–1954 who were offered teaching positions in New York City public elementary schools. A search was then made to find all schools in which two or more of these teachers were assigned to teach Grades 3, 4, 5, or 6. Of the 75 teachers who met these criteria, 19 were eliminated because the school was inaccessible, the principal did not wish to cooperate, or one or more of the teachers failed to accept the position or did not wish to cooperate. Of the 56 teachers who began the study seven dropped out during the year. (When one teacher in a school with two participating teachers resigned, the other teacher in that school was perforce also dropped.) Thus the sample on which the study was based comprised 49 teachers in all. Forty-six of the 49 teachers were women; 23 were teaching Grade 3, 13 Grade 4, 9 Grade 5, and 4 Grade 6. The teachers were scattered among 19 schools in the boroughs of Manhattan, Brooklyn, the Bronx, and Queens.

Variables

The study employed five variables purporting to measure one or another aspect of teacher effectiveness, three measuring dimensions of classroom behavior, and a number of variables designed to control extraneous variation.

Measures of Teacher Effectiveness

Adjusted reading growth. To measure growth in reading ability, four subtests of the California Reading Test (Elementary) —Word Form, Word Recognition, Mean-

[1] Grateful acknowledgement is due to Blanche S. Cogan and Yo Yee Chacko for invaluable assistance in the analysis of the data of this study.

ing of Similarities, and Interpretation of Meanings were administered to each class at the beginning of the year, and an equivalent form of the reading test was administered at the end of the year. The final reading test raw score of each child was adjusted by covariance analysis for his initial reading test score and for his intelligence raw score based on an administration of the Non-Language Section of the California Test of Mental Maturity. The within-classes regression coefficient was used as the best estimator of the mental ability effect to be removed. The teacher whose class showed the highest mean adjusted final raw score—that is, the greatest average improvement in reading—was considered most effective in stimulating pupils to learn to read. The reliability of these adjusted reading growth scores was estimated to be .84 (Mitzel & Medley, 1957).

Growth in group problem solving skill. To measure growth in group problem solving skill the Russell Sage Social Relations Test (Damrin, 1954) was used. This test requires the class as a group (without the teacher) to devise and execute a plan for constructing a replica of a model shown them, using special interlocking blocks provided by the examiner. It was administered to each class at the beginning and end of the school year. The protocols were scored in the prescribed manner by D. Damrin. This scoring procedure does not yield a single quantitative measure of group problem solving skill but, rather, 14 discrete ratings which are meant to be used to locate each class in a two-dimensional plane. The 14 ratings obtained on the second administration of the test were combined using weights obtained by the method of reciprocal averages (Horst, 1935; Mosier, 1946). The composite scores lay along the principal common factor of the test.

The weights obtained in the second administration were used to combine the 14 scores from the first administration into a single composite. Scores on the second administration were predicted from scores on the first administration, and the deviation of the actual second score of a class from its predicted score was taken as a measure of the effectiveness of the teacher in improving the skill of the class in group problem solving. The reliability of these growth scores is not known, but scores on the first administration had an estimated reliability of .85 based on an internal consistency analysis.

Pupil–teacher rapport. Ratings of pupil-teacher rapport were obtained directly from pupils' reactions by administering the My Class Inventory (Medley & Klein, 1957) to each class at the end of the school year. The "halo" score derived from this inventory is based on pupils' responses to the following eight items:

Do you ever feel like staying away from school?
Do you like to be in this class?
Do you have much fun in this class?
Do you learn a lot in this class?
Are you proud to be in this class?
Do you always do your best in this class?
Do most of the pupils like the teacher?
Does the teacher help you enough?

The reliability of the teacher's score on this scale was .89 in the group of 49 teachers.

Teachers' self-ratings. At the end of the school year each teacher was asked to judge how well she had played each of three roles a teacher must play. In Role I, the teacher "is responsible for providing learning experiences which will result in students' acquisition of fundamental knowledge." In Role II, the teacher "is responsible for providing children with learning experiences . . . leading to good citizenship, personal satisfaction, and self-understanding." In Role III, the teacher is "a professional colleague of other teachers, supervisors, and administrators." The teacher was asked to indicate where she would stand in a typical sample of 20 teachers with respect to her effectiveness

in playing each of these roles. Only the self-rating on Role I (teaching fundamentals) was used in this analysis.

Principals' ratings. The principal or the assistant principal responsible for supervising each teacher—or both, when available—was asked to evaluate each teacher in the manner described above on each of the same three roles. An analysis of variance failed to reveal any significant differences between ratings of the same teacher on different roles, so all ratings of each teacher were pooled. This pooled rating was regarded as a rating of over-all effectiveness of the teacher by her supervisor or supervisors, and had a reliability coefficient of .89.

Dimensions of Classroom Behavior

Classroom behavior was measured with the OScAR technique (Medley & Mitzel, 1958). Six observers each visited all 49 teachers twice at different times over a three-month period, and made an objective record of behaviors for half an hour on each visit. Scores on three orthogonal dimensions were developed by factor analysis of these records.

Emotional climate (reliability .90) refers to the amount of hostility observable in a classroom; a high score indicates a room in which external manifestations of warmth and friendliness are common and hostile reactions rare. Behaviors typical of the classroom in which Emotional Climate is *high* are:

Teacher calls pupil "dear," etc.
Teacher demonstrates affection for pupil
Pupil demonstrates affection for teacher
Teacher makes pupil-supportive statement or gesture

Behaviors typical of the classroom in which Climate was *low* include:

Teacher uses sarcasm
Teacher makes reproving remark
Teacher frowns, glares, etc.
Pupil ignores teacher's questions
Pupil scuffles or fights
Pupil whispers

Verbal emphasis (reliability .77) indicates the degree to which verbal activities predominate. Behaviors typical of a class with high Verbal Emphasis include:

Pupil reads or studies at his seat
Pupil writes or manipulates object at his seat
Pupil (or teacher) uses textbook or workbook
Pupil (or teacher) uses supplementary reading matter
Pupil (or teacher) uses writing materials

In addition, such a class was observed to be having a reading lesson relatively often and a social studies lesson infrequently.

Social organization (reliability .83) has to do with the amount of social grouping and pupil autonomy in a class. A class scoring high was one in which it was relatively common to find the class broken up into two or more groups working independently, and in which the teacher talked relatively little.

Control Measures

It is well known, and often pointed out, that pupil learning is a function of many variables of which the skill of the teacher is only one. If the part played by the teacher is to be detected, it is essential that as many as possible of these other influences be controlled. Most of these differences are apparent either as differences between schools and communities or as differences between classes within the same school.

Class differences within schools. As partial controls on the characteristics of the class taught by each teacher, the class mean scores on the Non-Language Section of the California Test of Mental Maturity and on the initial administration of the Russell Sage Social Relations Test were used. It was felt that information about the average mental maturity of a class would be valuable in studying relationships between behavior and effectiveness. The individual slow student, for example, would have a very different school experience with a given teacher in a class

of bright pupils than he would have had with the same teacher in a class all as slow as he is. Therefore, in addition to the control on individual ability incorporated in the growth measure by the covariance adjustment, these controls on the average ability of the class seemed important.

Similarly, the initial score of a class on the Russell Sage Social Relations Test was felt to contain information about the classroom environment which would be helpful in explaining the behavior of the class. A class able at the beginning of the year to discipline itself well enough to devise a feasible plan for constructing the block model and to carry it out without the teacher's supervision presents a very different problem to the teacher than is presented by a class which disintegrates whenever the teacher is not there to discipline it.

A third control on differences between classes in the same school was obtained from the grade level. The grade number of the class—three, four, five, or six—was simply inserted in each multiple regression equation to remove part at least of any differences among classes that might be related to grade level.

Differences between schools and neighborhoods. The effects of the type of school and community in which a beginning teacher finds herself can hardly be ignored in any attempt to relate her success in teaching to her own and her pupils' behavior or to other characteristics of the teacher. Such influences were controlled in this study by analysis of covariance applied in a manner similar to that described by Kendall (1948). All variances and covariances were computed separately between and within schools, and only correlations estimated from covariation *within* schools were used in the study. Thus, since all comparisons were made between teachers and pupils within the same schools, all effects of school and neighborhood differences were effectively removed.[2]

TABLE 1

WITHIN-SCHOOLS INTERCORRELATIONS AMONG FIVE MEASURES OF TEACHER EFFECTIVENESS

	1	2	3	4	5
1. Average Adjusted Reading Growth		+.063	−.002	+.133	+.405*
2. Average Growth in Group Problem-Solving Skill			+.250	+.067	+.179
3. Pupil-Teacher Rapport				+.378*	−.054
4. Supervisors' Ratings					+.123
5. Teacher's Self-Rating (Teaching Fundamentals)					

* Significant, .05 level (29 *df*).

ANALYSIS OF DATA

The five variables intended to measure teacher effectiveness were intercorrelated with the results shown in Table 1. The results suggest that two aspects of effectiveness are being measured by these five sets of scores. One set seems to be related to the teacher's ability to teach reading and the other to her ability to establish good rapport with pupils. Teacher self-ratings on the teaching of fundamentals appear to reflect the former ability; supervisors' ratings the latter.

This apparent finding that the two kinds of ratings of effectiveness measure different aspects of teacher effectiveness suggested further analysis of the data. Each rating was regressed on the three direct measures of effects on pupils and on the three control variables; the results, after control variables which did not contribute to the prediction were removed, are shown in Table 2.

Supervisors' ratings had a multiple correlation of .44 with the five variables used. The only appreciable beta weight was that with pupil-teacher rapport. Teachers' self-ratings on the teaching of fundamentals,

[2] Examination of between-schools variation indicated that considerable distortion of the findings was removed by this technique.

TABLE 2

RELATIONSHIPS BETWEEN RATINGS OF TEACHER EFFECTIVENESS AND OTHER MEASURES OF TEACHER EFFECTIVENESS

Ratings of Effectiveness	Multiple Correlation	Beta Weights					
		Measures of Effectiveness			Control Variables		
		Average Adjusted Reading Growth	Growth in Group Problem Solving Skill	Pupil-teacher Rapport	Average Mental Maturity	Grade Level	Initial Group Problem Solving Skill
By Supervisors	.44	+.10	−.08	+.39	+.19	not used	−.19
By Self on Teaching Fundamentals	.45	+.39	+.17	−.09	+.03	not used	not used

TABLE 3

RELATIONSHIPS BETWEEN TEACHER EFFECTIVENESS AND CLASSROOM BEHAVIOR

Criterion of Effectiveness	Multiple Correlation	Beta Weights					
		Classroom Behavior Dimensions			Control Variables		
		Emotional Climate	Verbal Emphasis	Social Organization	Average Mental Maturity	Grade Level	Initial Group Problem Solving Skill
Average Adjusted Reading Growth	.55**	+.20	+.09	0	not used	+.52	not used
Growth in Group Problem Solving Skill	.26	+.06	−.09	+.09	+.24	not used	−.26
Pupil-Teacher Rapport	.49	+.32	+.28	+.09	not used	−.24	not used
Supervisors' Ratings	.56*	+.52	−.01	+.10	+.20	not used	−.32
Teacher's Self-Rating	.48	+.10	0	−.44	+.28	not used	not used

* $p < .10$, $df = 25$.
** $p < .05$, $df = 26$.

with a multiple correlation of .45, appeared chiefly determined by pupil growth. Clearly these five measures tap two aspects of teacher effectiveness—one related to pupil learning, the other to pupil morale.

Relationship of Effectiveness to Behavior

In order to answer the main question—about the relationship between classroom behavior and teacher effectiveness—the multiple regression technique was again employed. Six independent variables, the three dimensions of behavior and the three controls on class differences, were employed, with only the control variables which actually functioned in a given equation being retained. School differences were again controlled by the within and between procedure. Each of the five measures of effectiveness was used in turn as a dependent variable. The results appear in Table 3.

Measured growth of pupils in reading ability based on raw score increments shows very little relationship to any of the three dimensions of behavior, but seems to depend only on the grade level

in question; apparently (when allowances for individual differences in ability are made) there is a tendency for growth in reading (as measured in this study) to increase from grade to grade over this range. These findings suggest that a generalized maturational factor tends to outweigh other factors in the assessment of reading growth in the elementary schools. We recommend investigators of this problem restrict their studies to a single elementary grade at a time in order to provide a local control of this maturational factor.

Growth of pupils in group problem solving skill, as measured by improvement in Russell Sage Social Relations Test scores, seems not to be related to any appreciable degree either to recorded classroom behavior or to differences among classes within the same school.

Pupil-teacher rapport seems most closely related to Emotional Climate, rapport (naturally enough) being highest where Emotional Climate is warmest. There is a suggestion that rapport is likely to be better when the emphasis on verbal activities is above average, and a hint that it is lower at the upper elementary grade levels.

Supervisory ratings are related appreciably only to Emotional Climate among the behavior dimensions. Apparently a supervisor thinks that the teacher whose class is friendly and orderly is an effective teacher. There is a small negative weight on initial social relations scores.

Teachers who rate themselves as highly effective in teaching fundamentals tend to allow their pupils less opportunity to work in small, independent groups than those who rate themselves as less effective. Their classes seem likely to be slightly higher in average mental ability.

Discussion

Previous research in the measurement of teacher effectiveness has tended to indicate that supervisory ratings of teacher effectiveness and measures of how much pupils are learning from the teacher have little in common. Typical conclusions drawn from such studies are:

> Teacher rating scales ... are only slightly related to the observed pupil growth (Hellfritsch, 1945).
>
> ... evaluations based on ... supervisors' ratings and those based on measures of pupil growth and achievement were not significantly correlated (Anderson, 1954).
>
> ... supervisory ratings here provided are invalid [as measures of pupil gain] (La Duke, 1945).
>
> ... supervisory ratings ... seem to lack reliability and validity [as measures of pupil gain] (Jayne, 1945).
>
> The criterion of pupil change apparently measures something different than that measured by teacher ratings (Gotham, 1945).
>
> The three criteria ... [pupil gain, pupil evaluations, and a composite of five supervisory ratings] are not related to a greater degree than can be attributed to chance (Lins, 1946).
>
> Whatever pupil gain measures in relation to teaching ability it is not that emphasized in supervisory ratings (Jones, 1946).
>
> Employers' ratings of teaching ability are not related to pupil gains in information (Brookover, 1945).

The conclusions cited above are consistent with the zero order correlations between supervisors' ratings and the two measures of growth reported in Table 1, +.133 and +.067. The results of the present study further suggest that supervisory ratings do not correlate with growth because they reflect the emotional climate of the class, rather than how much the pupils are learning. Perhaps it is a bit unreasonable to expect a supervisor to tell how much a class is learning just by looking at it. The notion that he can do so seems to be based on two assumptions: that there is a pattern (or set of patterns) of behavior exhibited whenever optimum pupil learning takes place, and that the supervisor can recognize this kind of behavior when he sees it.

The data about relationships between classroom behavior and pupil growth reported in this study do not support the

first assumption, at least for the types of growth measured here. The inductive way in which the three behavior dimensions used were obtained (Medley & Mitzel, 1958) indicates that they probably exhaust the more obvious behavioral aspects in which classrooms seem to differ. If there are uniform ways in which teachers and pupils behave whenever the pupils are growing in reading skill, they are not readily apparent to reasonably sophisticated classroom visitors. Raters of teacher effectiveness must seek subtler cues than these. There is no indication here of what these cues may be.

Unfortunately, so few attempts have been made to relate teacher behavior to pupil learning (cf. Jayne, 1945; Morsh, Burgess, & Smith, 1956) that it is not clear whether supervisory ratings fail to relate to pupil learning because it is impossible to assess learning in process, or because supervisors in general do not know how—do not know what to look for. These findings call into question the relevance of the whole of a considerable body of research in teacher effectiveness which has used ratings of some kind as a criterion of teacher effectiveness.

The finding that teachers are fair judges of their own effectiveness in teaching pupils to read was also reported by McCall (1952).

All of these results, tentative as they are, raise more questions than they answer, and point up the importance—the urgency—of more research along the lines laid down in the report of the AERA committee (cited above), research which defines and measures dimensions of effectiveness and of classroom behavior and relates the latter to the former.

The problem of relating behavior of teachers to effects on pupils is crucial not only to further research in teacher effectiveness, but to the future of teacher education itself. If the main objective of the professional part of teacher education is to teach teachers how to teach, it is highly desirable (to say the least) that clear-cut research evidence be obtained showing how the teacher must teach in order to bring about optimum pupil growth, and that such findings be made a part of every teacher's preparation. The amount of research, completed or underway, which can yield such evidence is, to repeat, astonishingly small.

It may appear that a rather extensive network of inferences about teacher effectiveness has been built from this study on a small amount of data—after all, only 49 teachers were involved. It should, therefore, be pointed out that the conclusions drawn are tentative, and are of interest mainly as illustrating what useful and far-reaching findings could be realized in more extensive studies along similar lines, studies employing criteria of effectiveness derived from measures of pupil growth.

Summary

Five measures of effectiveness and measures of three dimensions of classroom behavior were obtained on 49 beginning teachers in New York City public elementary schools, and analyzed with statistical controls on differences between schools and differences between classes within schools. The five measures of effectiveness were found to center around two distinct aspects of effectiveness. Supervisory ratings and pupils' reactions to their teachers appeared to reflect the teacher's ability to get along with children; teachers' self-ratings and measures of pupil gains (in reading and social skill) appeared to reflect effectiveness in stimulating pupils to learn to read.

An attempt was made to find out what kind of classroom behaviors were associated with each type of effectiveness. Neither measured gains in reading nor gains in group problem solving skill were found to be related to recorded classroom behaviors of teachers and pupils. Pupil-teacher rapport was found to be related to emotional climate and, probably, to

verbal emphasis in classroom behavior. Supervisors rated those teachers who had the friendliest classrooms as most effective. Teachers who rated themselves most effective in teaching fundamental skills tended to allow their pupils less opportunity to work in small, autonomous social groups.

REFERENCES

ANDERSON, H. M. A study of certain criteria of teacher effectiveness. *J. exp. Educ.*, 1954, **23**, 41–71.

BARR, A. S., BECHDOLT, B. V., COXE, W. W., GAGE, N. L., ORLEANS, J. S., REMMERS, H. H., & RYANS, D. G. Report of the Committee on Criteria of Teacher Effectiveness. *Rev. educ. Res.*, 1952, **22**, 238–263.

BROOKOVER, W. B. The relation of social factors to teaching ability. *J. exp. Educ.*, 1945, **13**, 191–205.

DAMRIN, DORA. The Russell Sage Social Relations Test: A measure of group problem-solving skills in elementary school children. *Proceedings of the Invitational Conference on Testing Problems*, 1954, 75–84.

GOTHAM, R. E. Personality and teaching efficiency. *J. exp. Educ.*, 1945, **14**, 157–165.

HELLFRITSCH, A. G. A factor analysis of teacher abilities. *J. exp. Educ.*, 1945, **14**, 166–199.

HORST, PAUL. Measuring complex attitudes. *J. soc. Psychol.*, 1935, **6**, 369–374.

JAYNE, C. D. A study of the relationship between teaching procedures and educational outcomes. *J. exp. Educ.*, 1945, **14**, 101–134.

JONES, R. D. The prediction of teaching efficiency from objective measures. *J. exp. Educ.*, 1946, **15**, 85–99.

KENDALL, M. G. *The advanced theory of statistics*. London: Griffin, 1948.

LADUKE, C. V. The measurement of teaching ability: Study No. 3. *J. exp. Educ.*, 1945, **14**, 75–100.

LINS, L. J. The prediction of teaching efficiency. *J. exp. Educ.*, 1946, **15**, 2–60.

MCCALL, A. *Measurement of teacher merit*. Raleigh, N.C.: State Superintendent of Public Instruction, 1952. Publ. No. 284.

MEDLEY, D. M., & KLEIN, ALIX A. Measuring classroom behavior with a pupil-reaction inventory. *Elem. sch. J.*, 1957, **57**, 315–319.

MEDLEY, D. M., & MITZEL, H. E. A technique for measuring classroom behavior. *J. educ. Psychol.*, 1958, **49**, 86–92.

MITZEL, H. E., & MEDLEY, D. M. Pupil growth in reading: An index of effective teaching. *J. educ. Psychol.*, 1957, **48**, 227–239.

MORSH, J. E., BURGESS, F. F., & SMITH, P. W. Student achievement as a measure of instructor effectiveness. *J. educ. Psychol.*, 1956, **47**, 79–88.

MOSIER, C. I. Machine methods in scaling by reciprocal averages. *Proceedings, Res. Forum, New York, IBM Corporation*, 1946, 35–39.

(Received June 24, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

THE DEVELOPMENT OF UNDERSTANDING IN ARITHMETIC BY A TEACHING MACHINE[1]

EVAN R. KEISLAR
University of California, Los Angeles

The use of teaching machines for the teaching of spelling and arithmetic combinations has already been shown to have merit (Skinner, 1954; Pressey, 1927). And studies have demonstrated that automated teaching can result in more than simple rote learning (Porter, 1957; Ferster and Sapon, 1958). In this study, the problem was to explore the possibility of using a multiple-choice method for the automated teaching of "understanding," specifically, an understanding of areas of rectangles. By *understanding* is meant the ability to answer a variety of questions different from those encountered during training but belonging to the same general class; the broader this class is, the greater is the understanding.

Essentially, the paper describes an attempt to devise a program for the teaching of understanding, together with the principles underlying its construction and weaknesses that were encountered in the use of the program. No information is available as to what a comparable group of pupils would have learned in regular classes. To have provided such information would have required a fairly large sample of teachers and classes. But more important is the fact that, at this writing, it is premature to compare the teaching machine approach with regular classroom instruction. Programs for use with such machines need improvement before such studies can have much meaning.

Apparatus

The teaching machine used in this study was an extensive adaptation of the Film Rater used by the Navy for teaching aircraft identification. Multiple-choice items on a Kodachrome strip-film were projected in sequence upon a viewing plate. The learner responded to each item by pressing one of five buttoms. If the answer was correct a green light was turned on and the next item could be brought into view by pressing a special button. But if this answer was wrong a red light came on; only after turning off this red light could the learner try again. To proceed to the next item the learner had to answer correctly.

A special device recorded a graph of all right and wrong answers for each item. If the wrong answer was given the pen moved to the right one-twentieth of an inch. For each correct answer the pen moved vertically an equal distance to a new line. Hence the subject's performance for any item could be read from the horizontal line on the graph corresponding to this item number.

The Item Set

The total program consisted of 120 items, 10 of which instructed the learner how to operate the machine and informed him of the goal to be attained. The remaining 110 items were constructed to provide a sequence beginning with concepts of squares, rectangles, length, and width. Following items requiring the pupil to indicate the number of square units in rectangles, the concept of area was presented. Applications included paint coverage, rug size, and tile laying, followed by practical problems of adding and subtracting areas and finding the length or width

[1] Appreciation is expressed to the staff of the University Elementary School of the University of California, Los Angeles, and to Mrs. Pauline Bart for their cooperation in this study.

of rectangles. The set concluded with items involving cost.[2]

Principles of Programming

Several of the principles outlined by Skinner (1958) and illustrated in his completion-item set were applied to this multiple-choice approach. To illustrate the principles discussed below, a short sequence from the program, Items 18 through 29, is presented in Fig. 1. The original program was in color.

1. The step from each item to the next in the sequence should be small enough so that the learner almost always gets each item right. Although Homme and Glaser (1958) and Coulson and Silberman (1959) found that smaller steps resulted in better learning and took less time per step than larger steps, definitive evidence on this issue has not yet been obtained. In this study it was assumed that if a pupil selected the wrong alternative to an item he did so either because he was improperly selected for the program or because of inadequate prior learning in the program itself. The programmer should make sure that a pupil learns enough before an item is presented so that generalization to this item is practically assured.

In opposition to this line of reasoning is the argument in favor of higher item difficulty that the learner is encouraged to "formulate his own hypotheses and try them out." This procedure may have merit if information is supplied to the student showing why the alternative is incorrect (Crowder, 1958) or if branching in the program permits special remedial instruction. But where, as in this study, the pupil is informed only that he is wrong when he makes an error, he is given no more information than when he gets the item correct. One fifth grade subject, after completing the program in this study, commented, "It's hard to know why you get something wrong. When I got it right, I knew it. When I got it wrong, I didn't know why." Since the absence of an explanation is likely to heighten the aversive consequences of failure, it appeared most desirable to adopt a minimum difficulty level for the items in this study. If it seemed helpful to emphasize that certain responses were wrong, instead of having the student learn this by "being wrong," special items were constructed for this purpose, e.g., "Which of the following figures is NOT a rectangle?"

2. Skinner's use of the vanishing stimulus and the prompt facilitates the occurrence of the correct response. For example, in Items 25 and 26 colored rows of squares encourage the right answer. In Item 27 these become merely dotted lines which vanish completely in Item 28. In Item 20 the correct answer "1 ft." is prompted by the dimension "1 ft." in the diagram.

3. To promote generalization within a broad class of items, a process necessary for understanding, the learner should acquire a variety of verbal responses which might later be used, through intraverbal associations, to evoke other appropriate responses. This is simply a process of mediated or secondary generalization, a complex example of which is found in Judd's theory of transfer through verbal principles. For instance, it was judged that if the pupil learned to group squares within a rectangle by rows, as in Items 25 through 28, he would acquire intraverbal associations which would promote the correct response in learning later in the program to multiply the length by the width to find the area and, still later, to divide the area by the length to find the width. In other words, with these items the pupil is being prepared to "gain in-

[2] This program, consisting of a Kodachrome strip film of 120 frames, has been deposited with the American Documentation Institute. Order Document No. 6080, from ADI Auxiliary Publication Service, Library of Congress, Washington 25, D. C., remitting in advance $2.00 for microfilm, $16.25 for 35-mm. enlargement prints. Make checks payable to Chief, Photoduplication Service, Library of Congress.

18. Squares can be large or small. But if each side on a square is one foot long, this is called a square foot. Which figure is one square foot?

A.

B.

C.

D.

E.

19. Here are two squares: They are exactly the same size. Suppose I put the squares together side by side like this.

What kind of a figure will the two squares together make if I leave out the middle line like this?

A. Another bigger square.

B. A rectangle.

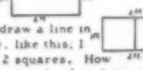
20. Here is a rectangle. It is 2 feet long and 1 foot wide.

Suppose I draw a line in the middle, like this. I will make 2 squares. How long will each side of each square be?

A. 1 ft.

B. 2 ft.

C. 3 ft.

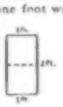
21. Suppose you have a rectangle which is 3 feet long and 1 foot wide like this 

How many squares, 1 ft. long on each side, can I make out of this rectangle?

A. One

B. Two

C. Three

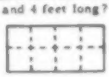
22. How many square feet are in a rectangle which is two feet high and one foot wide like this 

A. One

B. Two

C. Three

D. Six

23. How many square feet are in a rectangle which is 2 feet wide and 4 feet long? 

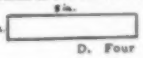
A. Two

B. Four

C. Six

D. Eight

24. Each side of this square is one inch long. This is called one square inch.

How many square inches are in this rectangle which is one inch wide and five inches long? 

A. One

B. Two

C. Three

D. Four

E. Five

25. Here is a rectangle which is 2 inches wide and 4 inches long. How many square inches are there in this rectangle?

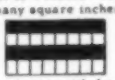
There are 2 rows with 4 squares in each row. How can we write the total number of square inches?

A. 2 + 2

B. 2 × 4

C. 2 + 4 + 2 + 4

D. 4 + 4

26. This rectangle is 4 inches wide and 7 inches long. How many square inches are in this rectangle? 

A. There are 4 rows with 4 squares in each row.

B. There are 8 rows with 4 squares in each row.

C. There are 4 rows with 7 squares in each row.

D. There are 4 rows with 14 squares in each row.

27. This rectangle is 5 inches long and 3 inches wide.

There are 3 rows of squares. In each row there are 5 square inches. How many square inches are there in the rectangle?

A. 5 + 5

B. 5 + 5 + 5

C. 5 + 5 + 5 + 5

28. Here is a rectangle which is 5 in. wide and 9 in. long. How many square inches are there in the rectangle?

A. 9 + 9 + 9 + 9 + 9

B. 5 + 5 + 5 + 5 + 5

C. 9 + 5 + 9 + 5

D. 5 + 5 + 9 + 9

29. A rectangle which is 5 in. wide and 9 in. long contains 9 + 9 + 9 + 9 + 9 square inches. What is a quick way of finding how many square inches this is?

A. Add 9 and 5.

B. Add 9 and 9 and 9 and 9 and 9.

C. Multiply 9 by 5.

D. Divide 9 by 5.

FIG. 1. A SAMPLE SEQUENCE OF ITEMS, NOS. 18–29, TO ILLUSTRATE PROGRAMMING PRINCIPLES

sight into why" you multiply in one case and divide in the other. Appropriate intraverbal associations, such as verbal principles, definitions, or characteristics, should function to extend the pupil's learning to entirely new items which involve the same principle or concept. If this indeed can be accomplished, the use of multiple-choice items in automated teaching results in something more than "mere recognition" of the right answer.

4. Procedures or concepts which are not otherwise involved in the sequence of items should be reviewed periodically. For example, a review of the process of grouping squares into rows, originally presented in Items 25–28, was provided later by six items occurring at intervals throughout the program.

5. Other techniques included the repetition of the correct answer on the succeeding item (as in Item 29), the irregular appearance of "interesting" colored pictures accompanying the item, and the use of a variety of forms of the multiple-choice item.

Subjects

Fourteen experimental *Ss* and 14 controls, individually matched on the basis of intelligence, sex, reading ability, and pretest scores, were selected from the fifth and low sixth grades. All *Ss* showed competence in multiplication and division but little acquaintance with the topic of area. A fifteenth *S* who had completed the program was dropped from the study because of an automobile accident prior to the posttest. The control *Ss* were given no special instruction of any kind. They were used to control for the effects of incidental learning such as that which might result from the administration of the pretest.

Pretest and Posttest

Both of these tests were of the free-answer or essay type. The pretest consisted of 12 problems involving multiplication and division, in addition to 8 problems dealing with areas of rectangles. The posttest consisted of the same eight problems of the pretest on area plus another eight most of which were more difficult. Sample problems on the posttest were:

1. A gallon of paint will cover an area of 200 sq. ft. How long a stretch of fence could you paint with 2 gallons if the fence is 5 feet high?
2. A sheet of cardboard is 3 feet long. It weighs exactly 4 ounces. What is the area of this cardboard if it is 2 feet wide?
3. A sheet of paper is 10 in. wide and 12 in. long. This sheet of paper is cut into strips. The strips are laid end to end and then joined with Scotch tape. This long strip now looks like this:

Can anyone tell what the area of this long strip is? .....
If so, what is it? .....

Procedure

Experimental *Ss* operated the machine for two or three periods on successive days. The total time spent with the machine ranged from one hour and 30 minutes to slightly over two hours. The posttest was given the day following the end of such machine instruction to each experimental *S* and his control.

Results

The mean score of the experimental group on the posttest was 12.4, with a standard deviation of 5.6. The corresponding control mean was 5.4, with a standard deviation of 3.7. Since the posttest scores for the experimental group showed a much greater variance than did the control, a sign test was used. All except one of the experimental *Ss* showed a higher posttest score than did their matched controls, a difference which is significant at the .01 level. Although the experimental group answered every item, except one, of the posttest better than the control group, most of these pupils missed several items

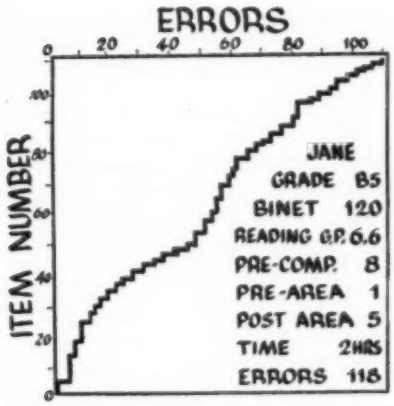


FIG. 2. GRAPHIC RECORD OF THE POOREST PERFORMANCE.

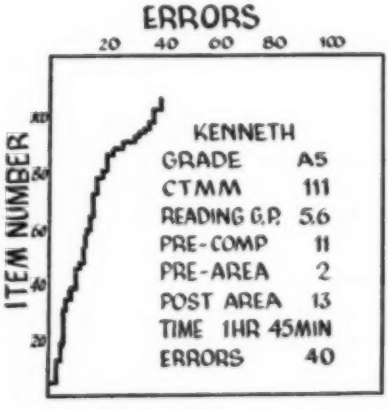


FIG. 3. GRAPHIC RECORD OF A TYPICAL PUPIL.

which were similar to those presented in the program; they learned far less than what had been expected.

The total errors on the program ranged from 3 to 118 with a mean of 54. Jane's record, presented in Fig. 2, was the poorest performance and indicates very little learning as measured by the posttest. Part of her problem may have been inadequate ability in computation, since her computation score was relatively low. Even for a

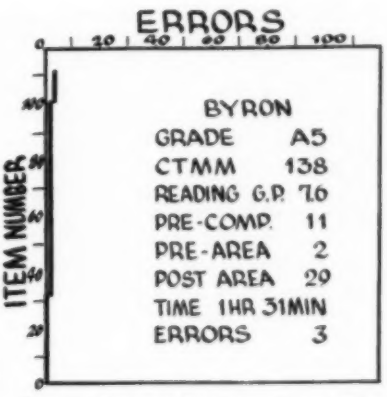


FIG. 4. GRAPHIC RECORD OF THE BEST PERFORMANCE.

typical student like Kenneth, whose record is shown in Fig. 3, the program was too difficult; 40 mistakes is entirely too many according to the criterion adopted. The program appeared to be ideal for Byron, whose record appears in Fig. 4. Although he showed little ability in the field of area on the pretest, Byron's performance on the posttest was outstanding. He was able to generalize from his training so well that on the posttest he solved completely new problems of obtaining the area of a parallelogram and a triangle.

The rank order correlation between total number of errors on the program and the gain on the posttest was $-.83$. While this of course supports the hypothesis that the optimum difficulty level of each item should be low, it does not permit, in itself, any such conclusion. The rank order correlation of mental age was .52 with gain on the posttest and $-.79$ with number of program errors. On the basis of this limited sample it appears that the program was more appropriate for the brighter children.

Desirable Revisions in the Program

In the absence of definitive evidence on the question, it appears that the major weakness of this program was that it was

too difficult for most of the pupils. Revisions should be made along the following lines:

1. Since the reading load was probably a major obstacle for many pupils, sentences should be shorter and the total amount of reading less for each item. Possibly several versions of the program, at different reading levels, would be desirable.

2. The steps in many if not most cases could be smaller. For example, Item 22, which was missed by three pupils, could be preceded by items analogous to 19 and 20. Item 24, also missed by three pupils, could be divided into two items, the first introducing the concept of square inch only. Item 25 could be rewritten as two or even three items.

The greatest misapplication of the principle of small steps occurred in the latter part of the program. For the group of 15 pupils who performed the item set, on the first 10 items there was a total of five errors on the first attempt (out of a possible 150). On succeeding sets of 10 items this number of first-attempt failures increased until it reached 66 for the last set of 10 items. The relatively high difficulty of the items in the latter part of the program appeared to result from the fact that these more complex items required a diversity of other understandings, abilities which were not tested for in the pretest. For example, although pupils were selected on the basis of their ability to divide, many failed to relate division to the process of successive subtraction. The program failed to provide adequate introduction or review of these concepts.

Faced with this type of problem a programmer can either use a more adequate pretest to provide for better selection or he can add the additional items as required. When a great variety of item sets in arithmetic become available, the problem largely disappears. The prerequisites to any item set can be stated in terms of the successful completion of previous item sets. Conversely, the record obtained for a pupil with one item set might be used diagnostically to indicate what the next set, remedial, optional, or otherwise, might be.

3. A wider variety of items should be used for each new process. For instance, Items 25–28 should be supplemented with items which provide verbal statements as alternatives and more familiar illustrations of the problem. As another example, on many of the posttest questions asking for an area, *Ss* wrote only the correct number failing to indicate the square units involved. Although six items had been presented in the program to reduce this kind of error, these items unfortunately were all stated in exactly the same way. The item stem called for an area and among the alternatives one distractor listed only the correct number. Although completion items may be necessary to teach this type of behavior, better results in this program could probably have been obtained if, instead of the single form, a variety of multiple-choice forms had been used, e.g., use of "none of these" as an alternative, asking "What is left out in this answer?" or asking for the appropriate rule.

Conclusion

The use of multiple-choice items in automated teaching appears to have some effectiveness under these conditions in teaching understanding, as herein defined, even though the criterion was a free-answer test. Although the average pupil did not show as high a degree of competence on the posttest as expected, the program was also far more difficult than intended. Since there appeared to be a strong relationship between success on the program and gains on the posttest, before the limitations and advantages of the multiple-choice method used in this study may be assessed, the program should be revised to include smaller steps and a greater variety of items. Whether the best performers on the present program would have learned more

from such a longer and simpler revision remains to be determined; to accommodate individual differences in ability two or three versions of the program may be desirable.

SUMMARY

Fourteen elementary school pupils responded individually to a set of 110 multiple-choice items in a teaching machine. The performance of each child was graphically recorded. Subjects performed significantly better on a test of understanding of areas of rectangles than did their matched controls who received no planned instruction on this topic. The principles of programming are discussed and illustrated. Suggestions given for the revision of the program, which appeared to be too difficult for most pupils, include the introduction of smaller steps and a greater variety of types of multiple-choice items.

REFERENCES

COULSON, J. E., & SILBERMAN, H. F. *Results of initial experiment in automated teaching.* Santa Monica, Calif.: System Development Corp., 1959.

CROWDER, N. A. Automatic tutoring by means of intrinsic programming. Paper read at the Air Force Office of Scientific Research and the University of Pennsylvania Conference on the Automated Teaching of Verbal and Symbolic Skills, Philadelphia, December, 1958.

FERSTER, C. B., & SAPON, S. M. An Application of recent developments in psychology to the teaching of German. *Harv. educ. Rev.*, 1958, **28**, 58–59.

HOMME, L., & GLASER, R. Relationships between the programmed textbook and teaching machines. Paper read at the Air Force Office of Scientific Research and the University of Pennsylvania Conference on the Automated Teaching of Verbal and Symbolic Skills, Philadelphia, December, 1958.

PORTER, D. A critical review of a portion of the literature on teaching devices. *Harv. educ. Rev.*, 1957, **27**, 126–147.

PRESSEY, S. L. A machine for automatic teaching of drill material. *Sch. Soc.*, 1927, **25**, 549–552.

SKINNER, B. F. Science of learning and the art of teaching. *Harv. educ. Rev.*, 1954, **24**, 86–97.

SKINNER, B. F. Teaching Machines. *Science*, 1958, **128**, 969–977.

(Received July 6, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

TRANSFER FROM CONTEXT BY SUBTHRESHOLD SUMMATION[1]

GEORGE M. HASLERUD
University of New Hampshire

Transfer needs a more adequate model. None of the three common paradigms of transfer, very well outlined by Vanderplas (1958)—the Osgood surface, generalization, and mediating association—has yet solved the crucial problem of similarity. Moreover, there is yet no evidence that these paradigms can be applied fruitfully to situations more complex than the isolated motor or verbal situations where they originated and have been tested. Is there a possibility that these models have concentrated on problems which may be only tangential to the phenomena of transfer occurring in such out-of-the laboratory contexts as allusion, analogy, and relevance?

The twin problems of similarity and meaning both turn upon how a new situation is perceived by the subject (*S*). As soon as the relationship is perceived or the point is seen as similar to something known before, then the remainder of the learning task is no longer transfer but a routine organization of the details and an attainment of skill enough to meet the criterion. In other words, the "transfer" could be expected to manifest itself only in the first meeting or two with a new situation, as it does in the Gestalt insight problems, but its influence would largely be masked when total trials to criterion is the comparative measure.

A perceptual model of transfer based on context would, therefore, seem a natural solution. But before such a model can be formulated, experimental tests of perceptual behavior in substantial contexts must be made.

Most studies of context have not gone beyond a single synonym as context for a test word (Cofer & Shepp, 1957), or the beginning of an associative chain (Russell & Jenkins, 1955), or pre- and post-alternative words for a picture exposed tachistoscopically (Lawrence & Coles, 1954), or a cluster of words in the preexposure field, meaningful or nonmeaningful, depending on the relation to the subsequent stimulus word (Cohen, Gofstein, & Casey, 1959). Some found interference from contexts like these, while others found facilitation, especially if the relationship was close (Cofer & Shepp, 1957). A special kind of context, familiarity of a word as evidenced by recognition of its definition, was found to aid perception of nine-letter words exposed for 40 ms. (Haslerud & Clark, 1957).

The studies of more complex context are very limited in number. In 1954, the experimenter reported at the annual meeting of the Japanese Psychological Association a pilot study of his half-story technique, in which he had the collaboration of Koji Sato and graduate students, M. Akita and T. Kambe, of Kyoto University.[2] Japanese kindergarten children transferred to the second half of an illustrated story meanings and inferences from the first half as proved by contrast with those who met only the second half. If the story experimentally reached an apparent ending after the first half, however, little was transferred to the unexpected second half the next day. The study was only suggestive of the possibilities because of such problems as keeping a straight learning approach to transfer from becoming merely rote. Searching for a bypass of the difficulties led to the idea of a direct perceptual test of transfer. Such a

[1] This research was supported by a grant from the Central University Research Fund of the University of New Hampshire.

[2] The extensive pilot experiments in 1954 were made possible by the Supplementary Educational Allowance provided by the United States Educational Commission in Japan (Fulbright).

study has been done by Miller (1956) with supraliminal exposure of .15 sec. He found that 25 sentences were more readily perceived tachistoscopically when the successive sentences told a story than when they were in random order. But the continuity must have been attenuated by this fragmentary parceling out of meaning, especially on a very simple story for adults. In fact, the best record was by an *S* who did better on the random presentation. Evidently a more complex, meaningful, and interesting context is needed if one wishes to explore the ephemeral, every-day transfer sets.

The purpose of this experiment is to measure the transfer effect of such a context upon a subsequent phrase presented subliminally. The first hypothesis to be tested is that the preceding context will facilitate the perception of the continuation phrase. The second hypothesis is that context which is relevant will facilitate, while that which is irrelevant will not.

Procedure

One way to meet the need for meaningful but short contexts, in sufficient number and homogeneous enough to allow statistical test, is to use the classical limerick which, unlike the Lear variety, has a last line different from the first. This material is of interest to most *S*s, usually has a high Thorndike frequency count, and while the last line has no words identical with those in the first four lines, it does have a continuity of meaningful (even if sometimes farfetched) relationships, rhythm, and rhyme. From Reed (1925), 18 unfamiliar limericks were chosen whose first lines begin:

A maiden at college, named Breeze; A menagerie came to our place; A thoughtful old man of Lahore; There was an old lady who said; There was an old bore of Torbay; A cynic of much savoir-faire; Some amateur players, most brave; There was a young lady of Ealing; A pony, renounced for his sauce; There was a young lady of Florence; There was a young lady named Bright; An epicure, dining at Crewe; There was a young lady of Kent; There was a young poet of Kew; There was a young man who said, "Damn!"; A goddess, capricious, is Fame; There was a young lady named Kate; There was a bad schoolboy who baited.

Each *S* had the characteristics of limericks recalled to his attention by preliminary analysis of the familiar limerick, "There was a young lady of Niger," and the process of anticipating the fifth line, by the unfamiliar, "There was a young girl of Navarre."

The following procedure introduced each limerick: The *S* was instructed to accentuate rhythm and rhyme as he read aloud twice the typed first four lines. Following the suggestion to the *S* that he anticipate the fifth line, the *E* said, "In the tachistoscope you may see the last part of the fifth line of the limerick, or it might be something else. But report anything you see, even if it seems incomplete." For presentation in the tachistoscope at 24 in., the last half-line (usually three words) was on a separate sheet ink-lettered one-eighth in. high at fixation level and to the right of center of the midline.

A Harvard-Dodge tachistoscope had its pre-exposure field set at .40 sec. and the exposure time for stimulus material determined for each *S*. The limens for seven preliminary three-word endings of sentences similar to the stimulus materials were obtained by increasing the exposure time by .01 sec. each trial until the *S* could report perfectly. Then three-fourths of his lowest time was set as the unvarying exposure interval for the experimental last half-lines, with trials repeated every 10 sec. The *S*s were not informed that all exposures were the same, and their inference that the time was being lengthened was fostered by *E*'s ostentatious adjustment of an irrelevant disc of the apparatus while he said, "Let's try this now." At this subliminal level, about five exposures at the 10-sec. separation are needed for unrelated three-word phrases to reach threshold level, and a facilitating factor has room to indicate a

quicker response. If the *S* did not give a perfect report in 10 trials, the *E* proceeded to the next limerick. If the three-fourths formula proved inadequate for an *S* as shown by wide variation from five repetitions for unrelated material, the constant was changed. Out of 80 *Ss*, two required lowering to .6 and two required raising to .9.

Although temporal and areal summation have a long history in neurology and physiology, this experimenter has been unable to find any previous perceptual studies of comparative stimulus thresholds which present an unvarying subliminal stimulus at equally spaced intervals until sufficient exposures permit a correct response. The method of this experiment was empirically worked out only to the point where it was sensitive, economical of time, and less fatiguing for *E* and *S* than more conventional psychophysical methods. The comparisons of frequencies of exposures for two conditions allow the testing of each of our two hypotheses, but a research is planned to study systematically the relation of thresholds reached by this kind of summation with those reached by progressive lengthening of exposure for successive trials.

The *Ss* were 80 college students divided so that 20 male and 20 female students were used to test each of the two hypotheses. Because of the hour to hour and a half required for the individual testing of each *S* and to improve the meeting of appointments, the *Ss* were paid for their cooperation.

TABLE 1

SUBLIMINAL EXPOSURES TO REACH THRESHOLD FOR FIFTH LAST HALF-LINE OF NINE LIMERICKS PRECEDED BY OWN FOUR LINES AND OF NINE PRECEDED BY NONE

(Each *S* was own control)

Group	*N*	$\bar{X}_{Lim.}$	$\bar{X}_{None}$	θ_{xdiff}	*t*	*p*	*Ss* in Expected Direction
M	20	4.28	5.61	.28	4.75	$<.001$	18
F	20	3.71	4.69	.23	4.26	$<.001$	18

Each *S* as his own control had nine limericks in one condition and nine in the other, e.g., nine relevant and nine irrelevant. As a further control, each limerick was presented in both conditions, e.g., 20 *Ss* had it in the relevant condition and 20 in the irrelevant. The same was true for the group that tested the influence of presence or absence of previous context. For the relevance-irrelevance study, the fifth lines were randomized for each *S* by a table of random numbers, but for presence-absence, the two conditions were alternated.

RESULTS

For each *S* the data consisted of the number of trials to attain a perfect response in each of the two conditions. Because each *S* was his own control, the difference between his scores includes the correlation factor for related samples. The significance of the mean difference score from zero difference was found. Table 1 indicates that the perception of the last half-line of a limerick required significantly fewer subliminal exposures when preceded by the first four lines than when alone. Furthermore, as can be seen from the right-hand column, 36 out of 40 *Ss* individually showed this same direction, with the other 4 either exhibiting no difference or only a slight difference in favor of the isolated line. Hypothesis I, that preceding meaningful context has a facilitating effect on subsequent perception, is thus well supported.

But not every context will give the perceptual advantage. As proposed in Hypothesis II, the preceding context must be relevant to the exposed half-line. In Table 2 one can see that relevant endings are perceived significantly faster than the irrelevant. Also the tendency was true for 32 of the 40 *Ss* individually, well beyond chance expectancy ($p<.001$).

Since the limericks were balanced so that half the group had, e.g., the relevant situation and the other half the irrelevant, a

comparison can be made for limericks just as has been done in Tables 1 and 2 for persons. The results are not given in detail in this paper except to say that for both male and female groups five-sixths or more of the limericks were in the expected direction.

There are no sex differences in Table 2 (relevance-irrelevance conditions), but in Table 1 it is apparent that the males were significantly ($p < .01$) slower than the females when there was no preceding context. The males in Table 1 did not differ when context was present from males in Table 2, where there was the similar relevant situation, but the differences between absence of context and irrelevance were significant at $p < .01$.

The ratios of the relevant score to the combination of the relevant plus irrelevant scores and of the score with context to the combination of context plus no-context both distribute themselves in a normal-like curve, with mean at .44 and SD of .03. The characteristics of *Ss* differing in these ratios have not been investigated, but they may give further clues to the nature of ability to transfer.

TABLE 2

SUBLIMINAL EXPOSURES TO REACH THRESHOLD FOR FIFTH LAST HALF-LINES OF LIMERICKS WITH NINE RELEVANT AND NINE IRRELEVANT PRIOR CONTEXTS OF FOUR LINES

(Each *S* own control)

Group	*N*	$\bar{X}_{Rel.}$	$\bar{X}_{Irrel.}$	$\bar{d}_{Diff}$	*t*	*p*	*Ss* in Expected Direction
M	20	3.75	4.68	.22	4.23	< .001	16
F	20	3.52	4.51	.23	4.31	< .001	16

DISCUSSION

Both hypotheses were supported at a high confidence level. Variations in the difficulty of the limericks and variations in the *Ss*' capacity to catch the meaning to be transferred were cancelled out as factors by making limericks and *Ss* their own controls. Also, the number of limericks and the number of *Ss* were sufficiently large to give consistent results and highly significant *p*'s. With such adequate controls for the testing of Hypothesis II, the conclusion, at least for this college population, that relevant context facilitates perception can be accepted.

While the major controls for testing Hypothesis I were the same as for Hypothesis II, the experimental setting for Hypothesis I precluded the warm-up from reading the first four lines of the limerick in the "None" situation. Examination of Tables 1 and 2 would tend to discount the importance of warm-up as an explanation of the difference in exposures between presence and absence of context. The female students give the same results in Hypothesis II where it was controlled as in Hypothesis I where it was not. While the males in the similar "limerick" and "relevant" situations in Hypotheses I and II, respectively, had a difference in the means which was far from significant even at the .05 level, they needed significantly more exposures ($p < .01$), when given no preceding context, than the males for Hypothesis II where the context was irrelevant. A possible explanation for the males' slow perception of the isolated stimulus phrase may be the need for more cues to be summated when language resources are restricted, as has often been reported for males compared to females (e.g., Haslerud & Clark, 1957).

CONCLUSION

What, then, can one infer is transferred from context to facilitate the perception of the last half-line of this experiment? The non-Lear limericks used do not have identical nor even similar words but only rhyme, rhythm, and meaning to connect them to the previous context. The pursuit of relevance is probably like the narrowing of alternatives from familiarity with a word (Haslerud & Clark, 1957) or the presentation of limited alternatives (Lawrence &

Coles, 1954) in its effect on the readiness to be stimulated in a particular way. The anticipative set established by meaningful context apparently makes fragmentary stimuli sufficient when they fit, as for a rhyme. That the set is a projective anticipation seems plausible when one considers the negative cases, e.g., where the *S* attains a goal and thereafter transfers little (Haslerud, 1950) or the similar case of the passive, noninvolved *S* (Haslerud & Meyers, 1958).

From the present experiment some of the broad outlines of a perceptual theory of transfer can be surmised. But to complete the foundation of the theory and to deduce its practical implications, much more work will have to be done.

Summary

An experimental base for a perceptual model of transfer requires testing of two hypotheses: that preceding meaningful context will facilitate perception of a continuation phrase and, secondly, that the context must be relevant. Eighty college students of both sexes were shown individually in a tachistoscope the last half-lines of 18 non-Lear limericks. The exposures were set at ¾ of each *S*'s stimulus threshold for similar material, with trials repeated every 10 sec. until a perfect report. The context was the first four lines of the limerick read aloud twice by the *S*. Each *S* acted as his own control for the two conditions needed to test one of the hypotheses, and each limerick was similarly balanced within each group for the two conditions. Forty *Ss* were used to test each hypothesis. Both hypotheses were supported at $p < .001$ in the groups and with 80%–90% of the individual *Ss* in the expected direction. The results point to a readiness for even subliminal cues when a projective anticipation has been established by relevant context.

REFERENCES

Cofer, C. N., & Shepp, B. E. Verbal context and perceptual recognition time. *Percept. mot. Skills*, 1957, **7**, 215–218.

Cohen, M. L., Gofstein, A. G., & Casey, T. M. The effect of a meaning set on visual recognition thresholds. Paper read at Eastern Psychol. Ass., April 1959).

Haslerud, G. M. Properties of bi-directional gradients at subgoals. *J. genet. Psychol.*, 1950, **29**, 67–76.

Haslerud, G. M., & Clark, R. E. On the redintegrative perception of words. *Amer. J. Psychol.*, 1957, **70**, 97–101.

Haslerud, G. M., & Meyers, Shirley. The transfer value of given and individually derived principles. *J. educ. Psychol.*, 1958, **49**, 293–298.

Lawrence, D. H., & Coles, G. R. Accuracy of recognition with alternatives before and after the stimulus. *J. exp. Psychol.*, 1954, **47**, 208–214.

Miller, E. E. Context in the perception of sentences. *Amer. J. Psychol.*, 1956, **69**, 653–654.

Reed, L. *The complete limerick book*. New York: Putnam, 1925.

Russell, W. A., & Jenkins, J. J. *U. Minnesota Tech. Rep.*, 1955, No. 15 (Contract NSONR-66216).

Vanderplas, J. M. Transfer of training and its relation to perceptual learning and recognition. *Psychol Rev.*, 1958, **65**, 375–385.

(Received April 8, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

ADULT AGE DIFFERENCE IN PERFORMANCE ON A VISUAL-SPATIAL TASK OF STIMULUS GENERALIZATION[1]

FRANKLYN N. ARNHOFF[2]

Mental Health Research Unit, Syracuse, N. Y.

Previous investigations of the learning and performance of elderly subjects (*Ss*) in a multitude of situations have tended to show that older persons do not learn or perform as well as younger *Ss*, although the effects of chronological age differ widely on different tasks (Welford, 1958). While decreases in accuracy and slowing of reaction time have been reported frequently for older *Ss* (Welford, 1958), the differences between divergent age groups may be quite small.

Watson (1954) hypothesized that stimulus generalization would be less in aged as contrasted to younger persons, basing his prediction on the concept that the older person narrows his range of interests and concentrates on essential activities, ignoring the rest. Discussing the unconscious changes that are the concomitants of aging, Weinberg (1956) generalized that the aging organism, with lessened psychic energy to deal with all stimuli, begins to exclude them. If such exclusion of stimuli occurs, then the performance of aged *Ss* would be more stimulus bound, with the expectation of decreased stimulus generalization. However, a study on stimulus generalization and age by Smudski and Braun (1956) did not demonstrate differences in the stimulus generalization performance between young and old *Ss*, findings which may be related to Welford's (1958) observation that, since "easy" tasks make little apparent demand upon organically based capacities or past experience, people of all ages would perform about equally well on such tasks.

The purpose of the present study was to contrast the performance of old and young *Ss* on a visual-spatial task of stimulus generalization, specifically hypothesizing that while the young and old would have similar patterns of stimulus generalization, the errors would be consistently fewer for the older than the younger *Ss*.

Experiential and practice factors have been emphasized in comparative age studies as it is assumed that the older person will be under some handicap in novel situations which interfere with prior learning (Ruch, 1934; Korchin and Basowitz, 1957). Because of these factors a learning task was selected to be relatively simple in its psychological and motor demands, to have relatively simple task instructions, to require relatively little memory, and to be apparently free from past learning or experience. While findings from such a specific task may be difficult to generalize to other and more practical learning situations, it may lead to greater understanding of basic differences between age groups.

PROCEDURE AND SUBJECTS

The task was essentially that of making choice reactions on an apparatus modified from that of Brown (1951) and used subsequently in other studies (Bilodeau, Brown, & Meryman, 1956; Arnhoff & Loy, 1957). On a 3′ × 14″ plywood panel, painted flat black, seven 4-watt, 6-volt lights were mounted horizontally, equally spaced 4 in. apart. A red neon pilot lamp, 2 in. above the center lamp served as a fixation point and as a ready signal. The pilot lamp was turned on in random order, 3 to 5 sec. before the lighting of a stimulus light. *S* sat 5 ft. away holding a reaction button in the preferred hand. Response latency was measured

[1] Paper read at the Interhospital Conference, New York State Department of Mental Hygiene, Syracuse, N. Y., April 1959.

[2] Grateful acknowledgement is due Irving Lorge and Isabel McCaffrey for their helpful suggestions and statistical advice, and to the nurses and members of the Wagon Wheel for their cooperation in serving as subjects.

to the nearest .01 sec. by means of a Standard Electric Timer which was wired to begin timing with the lighting of a stimulus light and to stop when *S* reacted by pushing the reaction button. The experimenter (*E*) sat behind the panel. By means of switches, stimulus lights could be turned on in predetermined order after the ready light.

After the *S* was comfortably seated at the proper distance, in a semidarkened room, instructions were read specifying that the experiment was to study individual speed and accuracy. Each person was told to react as quickly and accurately as possible to the lighting of the center light by pressing the button he was holding. Each *S* was informed that other lights on either side of the center light occasionally would be lighted but the *S* was not to respond to them. If an error were made, *S* was to continue to respond as quickly and accurately as possible to the successive lights.

Before reading the instructions, E lighted the ready light to ascertain if *S* could see the light without difficulty. After *S's* affirmative response, he was requested to fixate on the ready light while first Light 1, then Light 7, and finally both Lights 1 and 7 simultaneously were lighted to determine if *S* could see the lights without difficulty. From *S's* response, it was reasonable to assume that all *Ss* used in the study were able to see all stimulus lights.

First, *Ss* were given 25 training trials to the center light (#4) only. After the training series, without interruption, the test series began in which each of the three lights to the left and right of the center light was presented four times, making 24 test trials to lights other than the center light. The 24 test trials were interspersed among 59 additional trials to the center light (#4), in six different orders, each beginning with a different one of the six test lights. In this manner, each test light appeared as the first test light after the training trials, an equal number of times. On each order, 9 old and 10 young *Ss* were tested, with *Ss* assigned to the different orders of presentation in successive order of their availability.

Stimulus generalization is defined operationally in terms of the number of responses made to lights other than the training light (center light). As response to the center light is the required response, responses to other lights are considered as errors. According to theory (Hovland, 1951), a gradient of generalization was expected, i.e., fewer errors to the lights progressively further from the center light.

The young *Ss* were 60 volunteer or practicing nurses. A local social club for older persons provided 54 volunteer older *Ss*. The older volunteers from a select, nonhospitalized sample of older persons probably supplied a more adequately functioning group of *Ss* than frequently is used in studies of this sort. The age range for the young group was 19–27 years, ($\bar{X}$ 20.4, σ 1.2), and for the old group, 60–82 years ($\bar{X}$ 70.1, σ 7.6). All the young group was female; the old group consisted of 22 males and 32 females. Since no sex differences could be demonstrated on the dependent variables of errors and reaction time, data for the male and female *Ss* in the old group were pooled.

No consideration of differences in education, socioeconomic class, etc. was given since indices of such variables appear to have different meanings when applied to the young and old (Arnhoff, 1955; Dennis, 1953). Verbal intelligence does not seem to be related to performance on this task (Arnhoff and Loy, 1957) and, therefore, was not measured.

Results

Analysis of Errors

Table 1 shows the percent of responses made to each light by each group. Except for a minor reversal at Light 3, the old group was in all instances less responsive than the young. The average of the differences between comparable percentages (light by light) was 4.0, a value which is significantly different from zero at the .05 level of confidence ($t = 3.01$, 5 df). Table 2 presents the percentage of *Ss* responding to each light at least once, and shows fewer *Ss* responding in the older group. In this instance the average of the differences between comparable percentages was 14.25, a value which is significantly different from zero at the .01 level of confidence ($t = 4.67$, 5 df).

Comparison of the two distributions of error scores (Table 3) yielded $\chi^2 = 13.1$, which, with 4 degrees of freedom, is significant at better than the .02 level of confidence (cells were combined where expected values were too small), indicating further the differences in responsivity between the two groups. Examination of the

error distributions for the two groups, however, reveals all *Ss* in the young group making at least one error, while 9 of the old *Ss* made no errors at all (zero error scores). The probability of obtaining such a difference in proportions (errors vs. errorless) by chance alone is less than .001 (Fisher's Exact Method).

Although the results thus far confirm the expectation of less responsivity (fewer errors and therefore by definition less stimulus generalization) in the old group, the unexpected and significant difference in the number of error-free performances suggested the desirability of further analyses of the data to examine and contrast the performance of the two groups *when error was made*, thus excluding from these analyses those old *Ss* who made no errors. Since studies using similar equipment (Mednick, 1955; Smudski & Braun, 1956) have reported few if any zero error scores, there was the possibility that these error-free *Ss* were atypical in other ways also. This approach to the data seemed to offer a safeguard against conclusions assumed valid for the majority, yet which might be markedly and disproportionately influenced by a few, possibly atypical *Ss*.

Omitting the 9 zero score *Ss* from the old group, the percentages were recomputed for the distribution of error scores (Table 3). The χ^2 between the two groups was now nonsignificant ($\chi^2 = 3.5$, 3 *df*, p .05, 7.82). While the general tendency towards less responsiveness (lower error scores) continues to be shown in the old group, the differences now are not significant. The major difference between the two groups of *Ss* is apparently the number of zero errors in the old group.

As can be seen from Table 1, the obtained generalization gradients are irregular for both the old and young groups and are not bilaterally symetrical as would be

TABLE 1

PERCENTAGE OF RESPONSES TO ALL LIGHTS

Subjects	*N*	Lights						
		1	2	3	4	5	6	7
Old	54	10.6	8.3	21.3	100	13.9	9.2	9.2
	45	12.8	10.0	25.6	100	16.7	11.1	11.1
Young	60	12.1	11.3	21.0	100	18.8	17.9	15.4

Note.—In the second row of data for the old group are the percentages based on omission of the 9 zero-error *Ss*.

TABLE 2

PERCENTAGE OF *Ss* RESPONDING TO EACH LIGHT AT LEAST ONCE

Subjects	*N*	Lights						
		1	2	3	4	5	6	7
Old	54	35.2	25.9	51.8	100	42.6	29.6	27.8
Young	60	40.0	36.7	68.3	100	51.7	55.0	46.7

TABLE 3

DISTRIBUTIONS OF FREQUENCY AND PERCENTAGE OCCURRENCE OF ERROR SCORES

Group		Error Scores															
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
Old	*F*	9	11	11	6	6	3	4	1	0	0	1	1	1	0	0	54
	%	16.6	20.3	20.3	11.1	11.1	5.6	7.4	1.9	0	0	1.9	1.9	1.9	0	0	100
	%	—	24.5	24.5	13.3	13.3	6.7	8.9	2.2	0	0	2.2	2.2	2.2	0	0	100
Young	*F*	—	11	9	13	10	7	1	2	3	1	2	0	0	0	1	60
	%	—	18.3	15.0	21.6	16.7	11.7	1.7	3.3	5.0	1.7	3.3	0	0	0	1.7	100

Note.—An error score is the total number of responses (errors) made by an *S* on all test trials.
The second row of percentages for the old group is based on data omitting the 9 *Ss* with zero-error scores.

TABLE 4

THE MEAN AND STANDARD DEVIATION OF INDIVIDUAL AVERAGE REACTION TIMES TO THE CENTER LIGHT AND FOR ERROR RESPONSES TO EACH LIGHT (IN HUNDREDTHS OF A SECOND)

		Lights						
		1	2	3	4	5	6	7
Mean RT	Old	34	32	36	41	36	37	34
	Young	30	27	27	31	27	30	27
Sigma RT	Old	19	20	20	18	21	16	16
	Young	9	12	9	8	8	10	8

theoretically expected. While these findings were unexpected, the reasons are not apparent from either the data or the procedures used. Comparable irregular gradients, however, have been previously reported (Mednick, 1955). Furthermore, while increased responsivity at the most peripheral light positions has been reported as early as 1939 (Humphreys), and more recently (Mednick: 1955, 1958), no adequate explanation has been offered.

While the differences in generalization between the two groups (Table 1) were found to be significantly different when based on the total group of old *Ss*, reanalyses of these data in terms of only those *Ss* who made errors markedly alters the magnitude of the differences (Table 1), so that comparison of the actual error performance of the two groups (average difference between percentages) was not significant ($t = 0.62$, 5 df).

Reaction Times[8]

Reaction times were analyzed for test trials to the center light, and for the other lights when an error was made. As reaction times are contingent upon response to a light, if no response was made the reaction time was infinite and counted as zero. The nine errorless old *Ss* were, therefore, necessarily excluded from analyses of error reaction time, reducing the old group N again to 45. They are, however, included in all analyses of reaction time to the center light as the percentage response here was in all instances 100%. For purposes of comparison, an average individual reaction time was computed for the 59 responses to the center light, and an average individual reaction time for error responses. The individual average reaction time for center light and for error then served as basic data for the age group comparisons.

The individual old and young reaction times were significantly different for the center light trials (F, $p < .05$), as well as for the error responses (F, $p < .01$). From Table 4 it can be seen that the average reaction times for the old group were consistently higher than for the young for the center light (Light 4) as well as for each of the peripheral light positions. Moreover, at each light position, the older *Ss* were more variable in average reaction times than were the younger *Ss*. As a further examination of the differences in the distributions of average reaction times for the young and old groups, they were dichotomized at reaction time 40 hundreths of a second. 100% of the young and 56% of the old had mean reaction times of 40 or less on the center light, and by the same dichotomy, 88% of the young and 71% of the old had average reaction times for error of 40 or less. Both differences in proportions are significant ($p < .05$), with the average reaction times for the older *Ss* consistently longer than for the young.

Relationship between Speed and Accuracy

The correlation between the individual total errors and the average reaction times for all trials (center light plus error responses) was $r = -.27$ for the old and $r = -.36$ for the young. While in both instances the correlations are significantly different from zero ($p < .05$), they are

[8] Although reaction times were measured in hundreths of a second, they are treated as scores, and the decimal point omitted in reporting.

not significantly different from each other, indicating that this relationship does not appear to be age related. While the negative correlation is to be expected, i.e., lower individual mean reaction time associated with higher total error scores, the magnitude of the relationship is small, indicating that those *Ss* in either group with the fastest reaction times are not necessarily those who make the most errors.[4] These relations, in conjunction with our other findings regarding errors and reaction times give support to Birren's (1955) observation that changes in reaction time may take place (within limits) without corresponding effects upon the number of errors made.

Reaction Times and Zero-Error Scores

As it was felt that the nine zero-error score *Ss* might be atypical in ways other than their error-free performance, their average reaction times were analyzed separately. All nine were above the median of the old group's average reaction times, and were scattered throughout this range (i.e., throughout the above-median range). The probability of all nine scores falling above the median is .002 (Fisher's Exact Method). The mean of their average reaction times was 46.3 (σ 6.9) as contrasted to the mean average reaction time of 39.9 (σ 12.6) for the other 45 *Ss* in the old group. As the error-free performances of these nine *Ss* are associated with longer reaction times, it is quite possible that slow reaction time was either a major determinant, or at least a major aspect of their error-free performances. In strict definition, an error-free performance might be considered as most stimulus bound and, therefore, deficient insofar as generalization is concerned. However, the reasons for their errorless performances of the nine old *Ss* are not evident from the data, with the possibility to consider, among others, that rather than a deficiency in generalization, the performance of these *Ss* may represent a basic difference in attitude or task orientation which results in their slowing down to maintain accuracy.

The slower response latencies of the older *Ss* is consistent with other findings (Welford, 1958) which show small but significant and consistent slowing in the response latencies of older persons. In view of the magnitude of the over-all relationship between speed and accuracy it is indeed difficult to generalize that slowing of speed compensates in older *Ss* to maintain accuracy. As shown previously (Welford, 1958), speed of reaction can be considered as a sum of two parts: motor speed and decision speed. Current opinion and evidence (Welford, 1958) suggest that central factors influencing decision speed are the more important. However, while neurological concepts are frequently invoked to explain age differences, it is likely that many of the observed phenomena are due to attitudinal differences towards experimental tasks and instructions, not only between divergent age groups but also within samples of persons homogeneous with respect to age. As in the present study, the errorless performance of the 9 old *Ss* may as well be due to conscious or unconscious attitudinal differences or, indeed, to both, as they are to neurological factors.

Although the results showing greater intersubject variability with increased age are consistent with previously reported work (Welford, 1958), it should be recalled that the time between the lighting of the ready light and the presentation of a stimulus was randomly varied between 3 to 5 sec. to reduce anticipatory responses. However, the shorter the interval (foreperiod) between ready signal and stimulus, the longer the reaction time (Gibson,

[4] The young group was too homogeneous in age to obtain relationships between age, and speed and accuracy. For the old group, although the number of cases at successive age intervals was small, no trend was observed.

1941; Mowrer, 1940), with differential effects due to age (Botwinick, Brinley, & Birren, 1957); that is, the shorter time span slows the subsequent reaction time of older *Ss* more than younger *Ss*. Consequently, the variable foreperiod in this study should result in greater variability of subsequent response latency in the old as contrasted to the young. Secondly, the younger group was much more homogeneous in age which may also result in less intersubject variability in this group. It is probable that some of the increased variability in the performance of older *Ss* in this and other studies may be related to methodological and sampling differentials.

Summary

Fifty-four aged *Ss* and 60 young *Ss* were examined on a visual-spatial task of stimulus generalization. The specific task was chosen because of its apparent lack of complexity, memory involvement, and relation to past experience—factors felt to handicap and lower the performance of aged persons. Data were analyzed on the basis of error performance and reaction time, as generalization is defined in terms of the number and position of errors.

Response latency was found to be longer for the older *Ss*, consistent with previous findings. Significant differences in generalization (number and position of errors) was found between the two groups, with less generalization in the old. However, nine old *Ss* made no errors at all in contrast to the fact that an error-free performance never occurred in the young group. To guard against disproportionate influence of errorless old *Ss*, data were reanalyzed, excluding these nine from the analyses. The differences in generalization between the two groups was now not significant. The average reaction times of these nine *Ss* were also found to be above the median for the old group.

A low, negative correlation between accuracy and reaction time was found, for both groups, with the lack of statistical significance between the two correlation coefficients indicating an apparent lack of age relatedness.

References

Arnhoff, F. N., & Loy, D. L. Relationship between two measures of stimulus generalization: Influence of intelligence upon performance. *Psychol. Rep.*, 1957, **3**, 465–470.

Arnhoff, F. N. Research problems in gerontology. *J. Geront.*, 1955, **10**, 452–456.

Bilodeau, E. A., Brown, J. S., & Meryman, J. J. The summation of generalized reactive tendencies. *J. exp. Psychol.*, 1956, **51**, 293–298.

Birren, J. E. Age changes in speed of simple responses and perception and their significance for complex behavior. In *Old age in the modern world: Report of the Third Congress of the Int. Assoc. of Geron.* London: Livingstone, 1955. Pp. 235–247.

Botwinick, J., Brinley, J. F., & Birren, J. E. Set in relation to age. *J. Geront.*, 1957, **12**, 300–305.

Brown, J. S., Bilodeau, E. A., & Baron, M. R. Bidirectional gradients in the strength of a generalized voluntary response to stimuli on a visual-spatial dimension. *J. exp. Psychol.* 1951, **41**, 52–62.

Dennis, W. Age and behavior: A survey of the literature. *USAF Sch. Aviat. Med., proj. Rep.*, 1953, Proj. No. 21-0202-0005, Rep. No. 1.

Fisher, R. A. Statistical methods for research workers. (11th ed) New York: Hafner, 1950.

Gibson, J. J. A critical review of the concept of set in contemporary experimental psychology. *Psychol. Bull.*, 1941, **38**, 781–817.

Hovland, C. I. Human learning and retention. In S. S. Stevens (Ed.), *Handbook of experimental Psychology*. New York: Wiley, 1951. Pp. 613–689.

Humphreys, L. G. Generalization as a function of method of reinforcement. *J. exp. Psychol.*, 1939, **25**, 361–372.

Korchin, S. J., & Basowitz, H. Age differences in verbal learning. *J. abnorm. soc. Psychol.*, 1957, **54**, 64–69.

Mowrer, O. H. Preparatory set (expectancy) —Some methods of measurement. *Psychol. Monogr.*, 1940, **52**(2, Whole No. 233).

MEDNICK, S. A. Distortions in the gradient of stimulus generalization related to cortical brain damage and schizophrenia. *J. abnorm. soc. Psychol.* 1955, **51**, 536–542.

MEDNICK, S. A. Stimulus generalization as a function of level of achievement imagery. *Psychol. Rep.*, 1958, **4**, 651–654.

RUCH, F. L. The differentiative effects of age upon human learning. *J. gen. Psychol.* 1934, **11**, 261–268.

SMUDSKI, J. F., & BRAUN, H. W. Changes in the gradient of visual stimulus generalization as a function of age. *Amer. Psychologist*, 1956, **11**, 374. (Abstract)

WATSON, R. I. The personality of the aged. A review. *J. Geront.*, 1954, **9**, 309–315.

WEINBERG, J. Personal and social adjustment. In J. E. Anderson (Ed.), *Psychological aspects of aging*. Washington, D. C.: American Psychological Association, 1956. Pp. 17–20.

WELFORD, A. T. *Aging and human skill*. Oxford: Oxford Univer. Press, 1958.

(Received February 9, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

SCHOLASTIC BEHAVIOR AND ORIENTATION TO COLLEGE

ROBERT C. BIRNEY AND MARC J. TAYLOR[1]
Amherst College

Soon after a student arrives at college two broad realms of activity begin to modify his behavior. On the one hand, he is required by the college to attend classes, study, prepare for examinations, etc. On the other hand, much of his time is spent with his peers whose norms regarding socializing, dating, etc. have a demand character. The particular effect of these two main types of activity will depend in part on the nature of the reinforcements which he finds within them. The curriculum offers grades, faculty recognition, prestige, and a sense of mastery of a certain body of material. The student culture offers prestige for social leadership, peer recognition, and a sense of escape from the curricular demands. Presumably, the student attempts to maximize his rewards and minimize his penalities, presumably by developing a mode of operation during his freshman year which, if moderately successful, probably becomes quite stable by the senior year.

This study was designed to permit identification of the various behavioral and attitudinal patterns reported by college seniors. The attitudinal pattern was conceived as an "orientation to college" to be measured by a specially constructed inventory of attitude items called the Orientation to College Inventory (OTCI). We have assumed that the scholastic (SCH) and social (SOC) orientations are measurably distinct. Thus, a student with a positive orientation toward social affairs (SOC) would seek most of his reinforcement within the realm of student social life, while the student with a positive orientation toward scholastic affairs (SCH) will seek his major rewards in curricular pursuits. Obviously, these two orientations interact, and it is expected that there are behavioral correlates of these patterns of interaction. The college behaviors under study were those of coping with the curriculum and choosing a career. Information was obtained by interviewing students at some length about the expectations they had held toward courses, the considerations generally adopted in choosing courses, and the reasons for taking the courses they did. Of course, it is impossible in this type of study to establish any causal relationships between the attitude and behavioral measures. Rather we must be content to establish the nature of the patterns which exist, with an eye to developmental studies later. Finally, the attitudinal and behavioral reports were supplemented with data gathered from the "public" records the students had established in extracurricular participation, counseling, and academic achievement.

The hypotheses tested in this investigation involved the relationship between our attitude measure, the Orientation to College Inventory (OTCI), and the reported behaviors. They are set forth later in table form after the measures and groups involved have been described in greater detail. These hypotheses emerged from the observations of the junior author regarding the differences between students in these behaviors. In treating the material on extracurricular activity and use of the college counseling service no hypotheses were stated. We simply wished to compare the major attitude-behavior pattern groups on these data.

PROCEDURE

The major results of this study are based on the data from a random sample of 57 Amherst College seniors. Every

[1] Now at Columbia Medical College.

fourth senior, who had been in continuous residence at the college since freshman year, was selected from the Student Directory. Data were obtained from all seniors selected except one who left school while the study was in progress. Also 63 freshmen were randomly selected for an administration of the OTCI. Twenty-four of these *Ss* volunteered to take the OTCI for a second time two months later.

The seniors were interviewed at the beginning of the second semester over a five-week period. All interviews were conducted by the junior author, 41 being conducted in the *Ss* fraternity house, and the remainder taking place in the laboratory. Interview time ran from one hour to two and one-half hours. The *Ss* responses were recorded in code on response sheets. Many interview items were presented on cards to the *Ss*.

Interview

The full interview schedule runs to twelve double-spaced pages and may be obtained by writing the senior author. The general interview procedure is best illustrated by the following example:

> The purpose of this interview is to find out how Amherst men choose courses. Anything you say, of course, will be kept in confidence. As you see I have assigned a code number to you. What you will say will not be directly associated with your name. (*E* enters code number on data sheet.) Here is a record of all the courses you have taken at Amherst. Are there any errors or omissions? Is your senior schedule in order? Let's go back to the time when you were choosing courses. Try to imagine how you felt at that time. (*E* selects a course that *S* took junior year and gives a brief explanation of what is meant.) I am going to ask you to tell me some of your expectations and feelings about these courses. (*E* hands *S* a loose leaf binder which contains 5″ by 8″ index cards on which the rating scales are typewritten.)
>
> The first card is:
>
> To what extent did you think it would be important to your general educational development?

1	2	3	4	5	6	7	8
I considered it extremely important to my general education			I thought it would be of moderate importance to my general education.			I didn't think it would have much to do with my general education.	

All of the course data were obtained using this interview technique. Some items dealing with Career Plans and the Major Field were informational, with a few being open-ended. In every instance the interviewer presented the item orally as well as visually, and clarification via standardized interviewer replies was possible. At the conclusion of the interview the *S* was given the OTCI to be filled out and returned.

Following the graduation of the *Ss*, records were obtained of extracurricular participation, contact with the college counselor, CEEB verbal scores and grades. At the end of the analysis of these data the senior author discussed the various *S* groupings with those college officials most likely to have impressions of the students.

The Orientation to College Inventory (OTCI)

Having postulated the attitudinal construct of orientation to college, we needed an instrument from which a student's orientation could be inferred. No existing measure seemed appropriate. This instrument, in addition to meeting ordinary psychometric requirements, had to be acceptable to *Ss* in the Amherst College (all male) population. Accordingly items were written with this population in mind and groups of students were asked to cooperate in the phrasing of the items. The final form consists of 60 Likert-type items taken from the original pool of 150 items. There are 19 SOC items, 20 SCH items, and 21 filler items.

Item analysis of the OTCI responses of the seniors was performed using a simple chi square technique in which the dis-

criminability of each item was tested against the sum of all others. First to be checked were five SCH items concerned with "grade-getting" since it was felt these might not correlate with the basic scholastic orientation of the other SCH items. They did not and were dropped. Eight additional items failed to relate to their respective dimensions and were eliminated from the final scores.

The scores used in this study are based on 14 SOC items and 12 SCH items, examples of which follow:

Social items

1. It is important to make a lot of friends at college. (A)
2. I believe that having a good time and getting a full share of fun out of college life is as important as any other aspect of your experience there. (A)

Scholastic items

1. I enjoy studying. (A)
2. I wish that I could take more courses than the college allows. (A)

Until cross-validation of these items on a larger senior sample is carried out, present findings must be interpreted with this limitation in mind.

The seniors showed Spearman-Brown coefficients of +.81 for the SCH and +.74 for the SOC, while the freshmen values were +.46 and +.40 respectively. The product-moment correlation between SCH and SOC scores was −.41 for the seniors and −.34 for the freshmen. Clearly, the freshmen gave much less consistent patterns of scores.

Subject Groupings

In taking the OTCI the senior sample gave four sets of responses, one for each of the college years. Correlations between successive years indicate little variability in these remembrances, though the freshman year scores do show the least relationship to the senior year scores. The SCH score for senior year yields a product-moment *r* with junior year of +.87, with sophomore year of +.67, and with freshman year of +.61. The SOC scores shows similar relationships. Senior-junior is +.88, senior-sophomore is +.86, and senior-freshman is +.66.

In the present analysis the *Ss* score was the mean of the "junior" and "senior" responses. The two distributions were split at the medians. Five *Ss* whose scores coincided with the SOC median were eliminated from subsequent analysis. On the basis of the two scores each *S* was assigned to one of four groups, High SCH-High SOC (A), High SCH-Low SOC (B), Low SCH-High SOC (C), and Low SCH-Low SOC (D). These *N*s are 10, 16, 15, and 11 respectively.

The hypotheses tested in this study, as they relate to the predicted behavior of these several groups, are presented in Table 1.

Results

Correlates of the OTCI

Available for all *Ss* were the College Entrance Examination Board Verbal scores, senior year scholastic averages, and cumulative averages through the first semester senior year. Table 2 presents the product-moment *r*'s for the OTCI scores and these grade measures, and includes the relationship between the two scores themselves. A previous study has led us to expect the CEEB verbal to relate to the cumulative average on the order of +.60. It appears that for seniors endorsement of the "scholastic" items is positively related to general verbal ability and actual academic performance. The relationship of the SCH score to cumulative average with verbal ability partialed out is +.29. These relationships were not found for the freshmen.

Comparison with Freshmen

The fact that the seniors completed the OTCI for each year of college as they remembered it invites comparison between the memory of freshmen orientations, and those obtained from the freshman class.

TABLE 1

HYPOTHESIZED RELATIONSHIPS BETWEEN OTCI SCORES AND INTERVIEW ITEMS

Expected Group Difference	Hypothesis	Interview Items
A + B (High SCH) greater than C + D (Low SCH)		Expected their courses to be more:
	1	1 important to their general education
	2	2 difficult compared to all the courses they could have taken
	3	3 enjoyable
	4	4 pertinent to their after college career plans
		Rate more important in choosing courses:
	5	1 the extent to which the course would contribute to (their) general education
	6	2 the pertinence of the courses to (their) future career plans
		Reasons given for choosing courses will show:
	7	1 a greater proportion of intrinsic reasons, e.g., ". . . really interested in it"
	8	2 a greater proportion of instrumental reasons, e.g., ". . . need it for graduate school"
	9	3 a lesser proportion of evasive reasons, e.g., ". . . it met at a convenient time"
C + D (low SCH) greater than A + B		Rate as more important in choosing courses:
	10	1 the difficulty of the type of material for you
	11	2 the grades that you thought you could obtain in the courses
C (Low SCH-High SOC) greater than A, B, D		Expect their courses to have:
	12	1 fewer hours of outside class work required
	13	2 to do fewer hours of outside class work
		Rate more important in choosing courses:
	14	1 The number of hours of work (they) thought they would have to do
B (High SCH-Low SOC) greater than A, C, D		Expect their courses to have:
	15	1 more hours of outside class work required
	16	2 to do more hours of outside class work
		Rate as unimportant in choosing courses:
	17	1 the number of hours of work they thought they would have to do

The freshmen scores show significantly more variability (nearly twice as much) than the seniors' "freshmen" scores, on both dimensions. The SOC means are quite close together, but the freshman SCH mean is five points higher, an untestable difference. The lower freshmen sample reliabilities, greater spread of score, and differences in mean suggest that freshmen may be considerably more confused about their orientation to college than are seniors.

TABLE 2

CORRELATES OF THE OTCI SCORES FOR SENIORS AND FRESHMEN

Variable	Seniors (N = 57)		Freshmen (N = 63)	
	CEEB Verbal	Cumulative Average	CEEB Verbal	Cumulative Average
SCH Score	+.37*	+.45*	+.10	+.08
SOC Score	−.12	−.20	−.05	−.02
CEEB Verbal		+.67*		+.58*

* $P < .01$.

Analysis of the Interview Data

The hypotheses were tested by use of the Mann-Whitney *U* test (Siegal, 1956). This required that the 52 *Ss* be placed in rank order for each item to be tested. Medians were obtained for each group, and the extent to which these medians depart by chance from each other in the predicted direction ascertained. Ranking was done by using the scale position the *S* endorsed, the number of hours he reported, or a classified proportion of reasons selected from a total number of reasons chosen. The exploratory analysis of informational items was done with the chi-square technique.

Test of Hypotheses

Table 3 summarizes the findings for the hypotheses summarized in Table 1. Of the 17 predicted relations only four fail to appear. These will be discussed subsequently.

Information was also gathered about the details of course selection practices. Table 4 contains the questions, parameter data and group difference levels of confidence for this portion of the interview schedule.

TABLE 3

TESTS OF SIGNIFICANCE FOR HYPOTHESES IN FIG. 1

Expected Group Difference	Hypothesis	χ^{2a}	H^b	*P*
A + B (High SCH)	1		18.4	.001
greater than C + D	2		3.9	.05
(Low SCH)	3		7.7	.01
	4		13.3	.001
	5	7.8		.01
	6	2.79		n s
	7		28.0	.001
	8		3.3	n s
C + D (Low SCH)	9	9.33		.01
greater than A + B	10	11.4		.001
(High SCH)	11		33.4	.001
C (Low SCH-High SOC),	12		3.6	n s
greater than A, B, D	13		3.5	n s
	14	4.47		.05
B (High SCH-Low SOC)	15		8.3	.01
greater than A, C, D	16		6.8	.01
	17	6.2		.01

[a] Median test (Siegal, 1956).
[b] Kruskal-Wallis (Siegal, 1956).

Five questions were asked concerning the major program. The groups do not differ in major field, time of major decision, majors considered, or change in major. The High SCH *Ss* report significantly greater confidence that they would choose the same major again.

Fifteen questions were asked about career choices and current plans. Here only the SCH dimension (Groups A and B) showed any power of differentiation. Eighty-one per cent of *Ss* with High SCH scores intended to pursue professional as opposed to business careers, while only 42% of the Lows reported professional aspirations. Since very few Lows planned to attend graduate business schools far more Highs would receive graduate training than Lows. This presumably reflects the ability and cumulative average relations presented. When asked how many changes of plans had taken place in the last six months, the High SCH *Ss* report significantly fewer than the Lows.

Finally, the interviewees were asked how much time they had spent studying for each course. These inquiries covered the courses taken during the junior year and the first semester of the senior year. In an effort to learn more about our groups we turned to the record of "public" life which they had left behind. This was done mainly to learn more about Group D, although we expected to discover something about the others as well. An analysis was done of the college records of extracurricular participation so as to examine the number and nature of different activities entered during the four years. Table 5 shows a comparison of the groups for study time, number of different activities over four years, and the nature of the activities.

It is clear the High SCH groups A and B had a record of wide participation while the Low SCH groups had a restricted

TABLE 4
GROUP DIFFERENCES IN COURSE SELECTION PRACTICES

Question	Range	Median	Group Difference[1]	*P* values
How many days did you take to consider the courses you selected? (From the time you picked up the course announcements until the time you handed in the program.)	1 to 30 days	3 days	A B (High SCH) more than C D (Low SCH)	$<.01$
How many hours did you spend talking with the following[2]				
friends	1 to 10 hrs.	3 hours	none	—
faculty	0 to 5 hrs.		B > A, C, D	$<.01$
Would you estimate the percent of your knowledge of courses that comes from:				
other students	10% to 98%	25%	C,D > A,D	$<.03$
college catalogue	1% to 90%	24%	none	—
faculty (by asking them)	0% to 75%	5%	A,B > C,D	$<.01$
your own observations	0% to 75%	24%	none	—
Have you experienced any conflicts while choosing courses which involved wanting to take more courses than you could?	Yes No 81% 10%		A,B,D > C	$<.001$
Involving a choice among courses that you didn't want to take?	Yes No 40% 60%		A,C,D > B	$<.01$
Have you sought to balance your courses in any way?	Yes No 71% 29%		C,D > A,B	$<.01$
To what extent has the hour at which the courses met influenced your choices?	Some None 37% 63%		C > A,B,D	$<.001$

[1] All *P* values obtained from chi square analysis of 4 by 2 tables.
[2] Deans, parent, and others elicited a response from only 3 *Ss*.

record. Turning to the nature of this participation we divided them into four groups: Service organizations, Talent organizations, Interest groups, and Sports.[3]

We see that the nature of the activity which the High SCH's displayed is spread over both Service and Talent organizations, while Group D can only be described as disinterested in all but Sports. However, it must be borne in mind that the Sports figure is based on participation in at least one varsity sport, and does not imply a sports "career." The fact is there were only 10 varsity letter winners in the entire sample, and they were evenly distributed across all groups. Furthermore, the Group D record for Sports is not different from the other groups.

To summarize this analysis of the public records of the students, it appears that the High SCH *Ss* were more apt to per-

[3] Service organizations were: Chest Drive, Blood Drive, Tutorial System, Mardi Gras and Prom Committees, Student Committee to the Faculty, and House Management Committee. Talent groups were: Band, radio station, yearbook, glee club, singing groups, theatre, literary magazine, and student newspaper. Interest groups were: Law club, Outing club, Philosophy club, Christian Association, and Sailing club. Sports was defined as participation in a varsity sport.

TABLE 5

COMPARISON OF GROUPS ON MEAN STUDY TIME, NUMBER OF ACTIVITIES, AND TYPE OF ACTIVITY

		Group A (N = 10)	Group B (N = 16)	Group C (N = 15)	Group D (N = 11)	P values
Mean study time (Hrs.)		20.9	28.8	16.9	18.3	.05 (F is 2.86, 3 df)
Number of activities	0-2	1	1	7	9	
	3-4	5	8	5	2	
	5-10	4	7	3	0	.001
Percentage taking part in one or more groups.	Service	40	69	33	27	.02 (B vs. A,C,D)
	Talent	60	75	33	27	.01 (A-B vs. C-D)
	Interest	30	44	60	09	.05 (C vs. A,B,D)
	Sports	70	56	73	82	.30 (A-B vs. C-D)

Chi square based on 3 × 2 table (A-B vs. C-D) is 17.24.

form public services, and exercise natural talent, while we know little more about Group D, the Low-Lows than we did at the outset. However, it seems safe to say that they were not using the usual campus organizations as sources of reinforcement. Even more interesting perhaps is the fact that Group C did not seem to be satisfying their SOC interests in campus-wide organizations. Presumably the fraternities served this purpose.

DISCUSSION

It appears that the response patterns on the OTCI have systematic relationships to the answers given to the interview questions concerning curricular behavior and concerns. Of course, it has not been established that the *Ss* did behave as they remember behaving, or have the expectations they now recall. But certainly the picture presented has a considerable amount of coherence, and if we are dealing here with the phenomenon known as "response set" these sets themselves show coherence. Furthermore some of the questions dealt with career plans, and the choices described seem consistent with the background information the *Ss* have given.

Of the two dimensions the SCH scale displays the most relationships. This was not wholly unexpected since the interview items are confined to scholastic behaviors. Presumably the introduction of questions concerning the social life and practices of the students would be required for the SOC dimension to generate relationships.

The two hypotheses which predicted the High SCH *Ss* would report choosing courses because they were instrumental to career or graduate work were not confirmed. Apparently all groups are apt to choose courses with career considerations in mind, since the other career item differences probably reflect ability differences rather than differences in orientation. These relationships suggest that a SOC orientation in college may carry no implications for the *Ss* orientation toward his work in graduate school or business. The other two hypotheses not sustained predicted that Group C would not only rate as important but also expect their courses to demand less outside work, and expect to do less outside work. Groups A and D are similar to Group C on these items. Since Group C did rate as important the amount of work demanded by

the course it may be that the distinction between amount of work and outside work was too refined for all but Group B who did expect more outside work, and considered the amount of work unimportant in choosing courses.

The patterns which have emerged are these. *Ss* with High SCH scores more often report they expected their courses to be important to their general education, difficult compared to all the courses they could have taken, enjoyable, and pertinent to their career plans. They rate as important considerations for choosing courses, the contribution of the course to their general education, the pertinence of the course to their career plans, and as unimportant the difficulty of the course material, and the grades they thought they would obtain. They give more intrinsic and fewer evasive reasons for choosing the courses they did. In choosing courses these *Ss* report spending more time, placing less reliance on students, being less concerned over "balance" in their schedule, and placing more reliance on the faculty for course information. Far more of these *Ss* are going to graduate school, and intend to enter professional as opposed to business careers. (Of the eight *Ss* who intend to enter teaching, seven had High SCH scores.) The positive correlation of the SCH with college average and CEEB verbal indicates that these men have the ability to do the work, and do it. It is very interesting that the pattern of correlations for freshmen, while reflecting the same direction of relationship shows no significant correlation between SCH scores and either ability or freshman year average. Obviously this may imply differences in sample, year in college, or a lack of reliability in the OTCI. Another possibility, suggested by the comparison of senior and freshmen scores is that orientation to college is something which emerges with experience as the individual discovers those sources of reinforcement which are available to a person with his ability and talent.

The groups patterns of the SCH-SOC scores show some signs of usefulness. In general the High SCH-High SOC Group A, seem most influenced by their SCH orientations, but the way in which the High SCH-Low SOC Group B departs from Group A is interesting. Group B reports expectations of more hours of outside class work, intention to do more outside work, and rate as unimportant the number of hours that they thought they would have to work. They spent more time with the faculty choosing courses, and had the least avoidance conflict in making choices. Furthermore, this group reported actually doing more outside study than the other three groups. Note that Group A did not join Group B in these particulars. Perhaps Group A's High SOC orientation blocked them from this kind of total committment to curricular concerns.

The Low SCH-High SOC Group C, distinguished themselves by rating as important the number of hours of outside work they thought they would have to do, by giving more consideration to the meeting time of a course, by reporting far less approach conflict in choosing courses, and although their reported study time was not significantly less than that of Groups A and D, their average of 16 hours per week seems, in some absolute sense quite low.

Group D, the Low SCH-Low SOC group was not uniquely distinguished on any of the criterion items. Of these *Ss*, 75% are below the college median for the CEEB verbal scores and this may be the key to understanding their position. Presumably they have to work for what they get, and this tends to limit their committment to the social orientation, while to claim a scholastic orientation would be patently unrealistic.

A check of the records of the college counselor showed that Group D was not distinguished from the other groups in number of counseling interviews for non-vocational purposes. Discussion with col-

lege officials also failed to reveal any distinguishing characteristics of Group D, unless it was the fact that these men were simply not well known. This was probably the most striking fact. As a group, the *Ss* with the Low SCH-Low SOC scores on the OTCI seem to have remained for four years "on the sidelines" of college life. Needless to say this type of finding continues to suggest additional hypotheses for future test.

We feel this study has accomplished its central purpose of establishing the nature of some of the various attitudinal-behavioral patterns which distinguish between college seniors. It appears that the chief variables at work are those of ability, talent, orientation to college, and reinforcement patterns provided by the college. At the end of four years these variables seem to constitute stable patterns of behavior having serious implications for an estimate of the educational experience of the student. At the moment we know nothing of the pattern of these variables early in the freshman year, or of the changes which occur by junior year. Since many of our interview items were of a reminiscence nature it will be necessary to compare them with the self-reported behavior of students at these stages in their college experience.

The present study suggests some hypotheses for these studies. Students who are most likely to display an extreme SOC orientation at the end of four years may be the middle ability group rather than the low ability group. Apparently these *Ss* are more apt to have scholastic aspirations which are destined to be frustrated. Another emergent hypothesis is that the nonscholastically oriented student may have a greater proportion of his concerns outside of college altogether. Finally, it is obvious that an intensive study of the social practices of student life is needed.

Summary

Fifty-seven Amherst College seniors were interviewed regarding past curricular expectations and practices. It was theorized that their curricular behavior would be systematically related to a two-dimensional conception of attitudes toward scholastic and social areas of activity. A sixty item Likert-type questionnaire dealing with college life was developed which yielded a Scholastic score and a Social score for each *S*. Seventeen hypotheses were tested and 13 confirmed. Exploratory relationships between scores and *Ss* verbal ability, grade point average, and participation in college groups were also determined.

Four coherent behavior patterns emerged when the scholastic and social scores were combined and divided into High SCH-High SOC, High SCH-Low SOC, Low SCH-High SOC, and Low SCH-Low SOC. Some implications of these patterns are discussed.

REFERENCE

SIEGEL, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

(Received April 28, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

THE RELATIONSHIP BETWEEN SALARY POLICIES AND TEACHER MORALE[1]

CLAUDE MATHIS

School of Education, Northwestern University

The topic of morale has long been of interest to psychologists. According to Haire (1954) probably no other field in the social psychology of industry can match the number of publications accounted for by studies of morale. Despite the generosity of research, morale still remains a variable which is difficult to define, although there is no doubt that the phenomena is real and does account for variations in behavior. The field of education has shown much interest in morale as it relates to job satisfaction in teaching, but this interest has resulted more in the voicing of opinions than in attempts to research the problem adequately.

Probably in no other area of education have opinions about teacher morale been as frequent and as diverse as those concerning merit rating as a method to be used in determining salary level. An abundance of opinion can be found in the educational literature today both for and against the desirability of pay differentials based upon professional evaluation of service rendered, with those individuals who oppose such evaluation, or merit rating, pointing out that teacher morale should suffer as a result of this procedure. Such arguments are based on the assumption that merit rating is a psychologically dangerous process since it invites an invidious comparison. Also, studies in the psychology of judgment demonstrate that when individuals are judging other individuals there is tremendous room for error and distortion in the perceptions involved. On the other hand, those who favor merit rating emphasize that it should increase the morale of the more competent individual, thus making him more productive, since his needs for achievement will be met more adequately by the recognition granted for outstanding service. The psychological efficiency of the total educational effort should be increased because merit rating would tend to discourage the less competent, thus helping to separate them from a profession for which they are ill suited. The importance to the educational psychologist of the teacher's morale is apparent, especially if morale is thought of as a psychological attribute which can be profoundly influenced by external factors such as the manner in which the teacher is paid for the services he performs.

PROBLEM

The specific purposes of this investigation were to design and test an attitude inventory for measuring teacher morale, and to determine if teacher morale is significantly related to salary policy in a small sample of school systems. A number of writers and organizations argue that high teacher morale and merit salary schedules are, to a considerable degree, mutually exclusive. If this argument is valid, one should find low teacher morale in school systems that use merit pay schedules. High, medium, or low morale might be found in school systems that use seniority type or single salary schedules. As a result of these opinions, the assumption was made in this study that a statistically demonstrable relationship exists between teacher morale in a school system and the use or nonuse of a merit type salary sched-

[1] The research reported here was done with the cooperation of B. J. Chandler, Professor of Education, Northwestern University, with the support of a research grant from the Graduate School of Northwestern University.

ule.[2] The assumption was also made that morale has identifiable behavioral dimensions that can be measured through the use of an adequately constructed instrument which would reflect a general consensus as to the meaning which the term "morale" has for education. From these assumptions the following hypotheses were derived for testing:

1. Morale differences exist between schools which include superior performance as a factor in determining pay and schools which use a single salary schedule based on seniority, educational background, and/or other factors.

2. Such differences in morale which exist are in the direction of lower morale in schools which use a merit pay plan.

The Morale Inventory

In order to test these hypotheses it was necessary to develop an attitude inventory capable of giving a quantitative index of morale relative to a person's role as a participating member of a school system. Morale should be reflected in the attitudes a person has about himself and about persons and things in his behavioral field. With respect to the research reported here, these attitudes center around the school environment.

In constructing the inventory, five attitude areas were identified which would provide a seemingly adequate measurement of the attitude possibilities inherent within the school environment:

1. *Self*—Attitudes about the self in relation to the role played in the school system.
   Sample item: I feel that I am an important part of this school system:
   ....With rare exceptions at all times
   ....Practically never
   ....Most of the time
   ....Only part of the time
2. *School*—Attitudes about the immediate aspects of the school situation such as working conditions, equipment, and the physical plant.
   Sample item: The size of the class (classes) I work with is:
   ....Large but satisfactory
   ....Completely unreasonable
   ....About right
   ....Unsatisfactory
3. *Community*—Attitudes about the community in which the school is located.
   Sample item: Getting acquainted with people in this community is:
   ....Almost impossible to accomplish
   ....A hard thing to do
   ....Done with some effort
   ....Quite easy to do
4. *Administration*—Attitudes about the way the school is administered, and attitudes about the people who administer.
   Sample item: When my teaching schedule is planned:
   ....I am never consulted
   ....My desires are always given preference whenever possible
   ....I am consulted and my desires made known but seldom considered
   ....My desires are usually considered
5. *Policy*—Attitudes concerning the policies and policy making functions relative to the school system.
   Sample item: Policies under which pay raises are granted are:
   ....Very unsatisfactory
   ....Sound and fair
   ....Unfair in many instances
   ....Reasonably satisfactory

Ten statements, of which samples are provided above, were included in the inventory for each area defined. These statements were by no means exhaustive, but appeared to represent a sampling of factors within each area which might involve morale. Three individuals who were familiar with morale studies in the field of psychology served as judges in the selection of the statements. A number of statements were rejected because, in the opinion of these judges, they were not ones which would have morale implications. The ten statements in each of the five areas represented the possibility of an attitude which

[2] For purposes of the research, the term "merit salary schedule" refers to any salary schedule for teachers which provides for salary above the regular schedule as a reward to teachers who are judged to be giving superior performance.

would correlate highly with feelings of morale. No statement was used which was not unanimously agreed upon by the judges.

The 50 statements which were derived were converted into multiple choice items in which *Ss* could choose one of four possible endings. The same expert judges who participated in the selection of the statements also participated in the construction of the possible choices. Endings were constructed for each statement which would indicate very high and very low morale. Two other endings were then constructed to indicate degrees of high or low morale, in between the two extremes. No attempt was made to determine whether or not the intervals represented approximately equal intervals except through the agreement of the judges. If the agreement was not reached on a particular ending, it was changed or revised until accepted by all three judges.

Before being printed in final form, the 50 statements were randomized, and the four possible endings were randomized for each statement. In this manner an attempt was made to avoid any pattern which might have acted to influence the choice of response. In order to obtain some indication of validity other than that involved in the agreement of the three judges who participated in the selection of the items, three individuals who knew nothing of the research project or the attitude inventory were asked to indicate which of the 50 items belonged to each of the five attitudinal areas. A contingency coefficient was computed for each of the three judges as an index of the degree to which they were able to assign statements to appropriate areas. These were positive and ranged from .76 to .79. The same individuals were also asked to rank the response endings in terms of the degree to which the endings indicated either high or low morale. The disagreement between judges, and the disagreement of each judge with the already predetermined direction of the endings was negligible.

The reliability of the attitude inventory was tested by means of the split-half method. The correlation between scores on the first 25 items with scores on the last 25 items yielded a coefficient of .74. Using this correlation in the Spearman-Brown formula for a double length test a reliability coefficient of .85 was obtained.

The Schools Studied

Ten suburban school systems were selected to participate in the study. Of the 10, two are large senior high schools, two are small senior high schools, and six are elementary schools. Two high schools and three elementary schools which use a merit salary system were selected first. Then an effort was made to match each of the five schools with a system using a single salary schedule. Factors considered in matching the schools included location in suburban communities with similar socio-economic population, number of teachers, true value of property, and current expenditure per pupil in average daily attendance. The matching represented the closest correspondence on all factors which could be obtained from a sample of possible schools in the immediate geographical area.

From the sample of the 10 schools involved, 614 inventories were administered and collected. The sample tested represented 336 *Ss* from merit salary schools and 278 *Ss* from single salary schools. With a few exceptions, the sample represented all the teachers employed in each school tested, with the number of inventories collected representing 100% of those administered. The inventories were scored by assigning a weight of 0 to the response ending which represented low morale, a weight of 1 to the next response interval above the low response category, a weight of 2 to the response interval below the high morale response, and a weight of 3 to the high morale response ending. An individual's score on the inventory would represent the addition of the assigned weights for the response endings selected by each

TABLE 1

MEANS AND STANDARD DEVIATIONS OF THE INVENTORY SCORES FOR EACH OF THE TEN SCHOOLS

Merit		Nonmerit	
Mean	*SD*	Mean	*SD*
114.82	13.45	109.31	12.72
113.90	14.83	112.68	14.27
107.95	13.96	118.68	13.74
113.68	10.55	115.40	11.79
119.19	13.30	113.91	13.07

individual. The instructions printed on the inventory emphasized the selection of only one response for each statement. The number of inventories in which items were skipped represented less than 2% of the sample.

RESULTS

The range of scores from the sample of 614 individuals in the 10 schools involved in the study was from a low of 66 to a high of 145. The mean of the total distribution of scores was 113 with a standard deviation of 13. Table 1 presents the means and standard deviations for each school of scores on the attitude inventories taken by teachers at that school.

When the total sample had been tested the data were tabulated in a manner that would allow for frequency totals with respect to schools and areas within the morale inventory. The data were then grouped so that a test could be made of differences between all 10 schools, as well as differences between schools grouped on the basis of pay plan. Also, the data, grouped on the basis of areas within the morale inventory, were examined to determine if any of the five areas produced a score significantly different from the other areas. The statistical procedure used was an analysis of variance technique.

The analysis of variance for the test data grouped on the basis of schools, regardless of the type pay plan used by the school, produced an F of 5.15, significant at the .01 level of confidence. The analysis of variance for the test data grouped on the basis of merit as compared to nonmerit schools produced an F of 2.03 which was not significant at either the .01 or .05 levels of confidence. The analyses of variance designed to test for differences in response between areas of attitude involved in the inventory produced Fs which were not significant.

From the results of these analyses of variance, it appears that the level of morale of personnel in the schools of the sample studied is not directly related to the type pay plan (merit or nonmerit) used in the school. Significant differences at the .01 level of confidence were found between schools relative to the manner in which personnel in the sample responded to the attitude inventory. The inventory seems to have been approached in a different manner from school to school with respect to the referential meaning it had for the *Ss*. If the inventory measures morale, one would expect differences between schools, since it is only logical to assume that morale conditions vary from one school to another. The sensitivity of the inventory is suggested by this difference between schools when one realizes that the schools used for the research have salary schedules well above the national average and quite similar to each other throughout the total salary range, whether the school happens to have a merit pay plan or a nonmerit pay plan.

The lack of significance between merit as compared to nonmerit schools in the sample suggests that level of morale does not differ over and above chance expectancy when comparing personnel in merit and nonmerit schools. If morale differences exist between schools in the sample, but not between schools classified on the basis of salary plan, then morale level should be the result of a multitude of factors in the school environment acting together with, or aside from, a merit or nonmerit type of salary policy.

No significant differences between areas involved in the inventory were found with respect to a comparison of the frequency of choice of response endings for questions in each area. Such a finding can be interpreted as meaning that individuals tended to approach the inventory as a whole, which could indicate that level of morale is a general factor reflected in a person's reaction to a school system as a whole. *Ss* used in this research responded consistently throughout the total inventory and did not reflect significantly different levels of morale relative to each area. Each attitudinal area appears to contribute in an approximately similar manner to the overall indication of morale given by the inventory.

While the results presented in this research are thought provoking with respect to a consideration of the merit pay issue in education, they should not be considered applicable to school situations other than the sample used here. The similarity of salary schedules in the sample reported here could be one reason for the lack of significant difference between merit and nonmerit schools. Perhaps there is little difference in level of morale because all the teachers are receiving an adequate salary regardless of the type pay plan.

One final word of caution should be stated. The inventory used in this research was designed as a measure of teacher morale. While necessary precautions were taken to insure its adequacy as a research instrument, it is subject to the same criticisms concerning reliability and validity which are common to most attitude inventories, one major criticism being that the inventory measures an attitude about an attitude, and not a direct behavioral response.

Summary

Personnel in 10 suburban school systems were given an attitude inventory designed to measure level of morale. The inventory was administered for the purpose of determining what differences in level of morale, if any, exist between schools which use a merit type salary schedule and schools which use a nonmerit type salary schedule. The statistical analysis of the data permit the following conclusions:

1. No significant difference in morale level was found between schools grouped on the basis of type of salary schedule.
2. A significant difference in level of morale, as measured by the attitude inventory, was found between the 10 schools involved in the sample.
3. No significant differences in indication of morale level were found between areas of the attitude inventory, suggesting that individuals within each school tended to approach the inventory as a whole rather than projecting differential feelings of morale into specific areas of the inventory.

REFERENCES

Coffman, W. E. Teacher morale and curriculum development: A statistical analysis of responses to a reaction inventory. *J. exp. Educ.*, 1951, **19**, 305–332.

Haire, Mason. Industrial social psychology. In Gardner Lindzey (Ed.), *Handbook of social psychology*. Cambridge, Mass.: Addison-Wesley, 1954. Pp. 1104–1123.

(Received May 13, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

DIMENSIONS OF LEADERSHIP BEHAVIOR IN CLASSROOM DISCUSSION GROUPS[1]

CLAUDE J. BARTLETT
George Peabody College for Teachers

Although many studies have been done in the search for the dimensions or traits of leadership behavior, an almost equally large number of dimensions of leadership have emerged. Certainly many of the dimensions found have been similar, and their difference may in some cases be a function of the interpretation of the studies. However, enough evidence has been accumulated (Stogdill, 1948) to indicate that leadership does not involve the same behavior in all situations.

The purpose of this study was to identify the dimensions of leadership behavior in classroom groups. It was conducted in a classroom situation where the group discussion method was used as the principal teaching method. This situation offered an excellent opportunity in which to study leadership behavior, since the class members worked in a discussion situation for approximately 12 weeks. Thus, the group members spent enough time in group discussion to demonstrate their abilities for leadership. In this particular situation, there was also considerable motivation to participate in discussion since a portion of the class grade is dependent upon contribution to group discussions.

METHOD

The subjects used for this study were students in an introductory course in educational psychology in which the small group discussion technique was used as the principal method of instruction. At the beginning of the quarter each class was divided into several groups of four to six students each. Each group worked separately on various topic questions and projects. The class procedure afforded an opportunity for the students to evidence leadership behavior and to become acquainted with other members of the group in a discussion situation.

Leadership was defined as being perceived as a leader by the other members of a discussion group. Students from three different classes (about 75 students) were asked to write essays describing the behavior of the most outstanding leader in their discussion group. They were asked to describe the behavior of the person chosen as a leader rather than write an essay on an ideal leader. Thus, the descriptions received tended to contain concrete behavioral descriptions of persons perceived as group leaders. These essays provided a source of short descriptions of the behavior which could be used as items in a descriptive check list. Three hundred phrases were obtained.

The Wherry-Winer Method (1953) of factor analyzing large numbers of items was utilized in analyzing the dimensions of leadership behavior. The items were grouped into the following categories for the purpose of analysis: (*a*) Intelligence (*b*) Personal acceptability (*c*) Motivation (*d*) Group organization (*e*) Democratic-autocratic (*f*) Self-organization. Items were selected to represent these categories on the basis of expert (advanced graduate students in psychology) agreement. A criterion of 80% agreement or better for the five experts used was required before an item was included in a category. At least 20 items were used to define each category.

The check list was administered to 100 students of the introductory educational psychology class, divided into three approximately equal groups. The first group

[1] This study was done at the Ohio State University as part of a doctoral dissertation under the supervision of Robert J. Wherry.

was asked to think of someone in the upper half of the class in terms of leadership qualities. The second group was asked to think of a person from the lower half in terms of leadership. The rest were asked to think of someone who was about average. After a student had called to mind a particular group member, he was asked to consider the 300 phrases, one at a time, as they applied to this individual responding to each phrase on a five-point scale of applicability.

A rating on over-all leadership by a group member was also obtained for each student. The applicability rating on each phrase was correlated with the leadership rating. Substantial positive correlations were obtained for almost all phrases, indicating that the phrases described the positive aspects of leadership.

Category scores were obtained for each of the 100 students on the six tentative categories. The six categories were intercorrelated and the *Z* portion of the Wherry test selection procedure (Stead & Shartle, 1940) was used to select categories. Categories whose unpredicted variance fell below .20 were eliminated. This procedure is suggested by Wherry and Winer (1953). Wherry and Winer also suggest that these categories may be assumed to be good estimates of the centroids that would be defined by the corresponding interitem correlations. Thus, the correlations between the selected categories and the 300 items may be considered as the oblique factor loadings as defined by these categories. A method suggested by Wherry (1959) was used to transfer these estimates of oblique loadings with simple structure preserved among the group factors. Two hand rotations were performed to clarify the simple structure and interpretation of the group factors.

Results

A general factor and four group factors emerged from the analysis. Almost all of the 300 phrases received substantial loadings on the general factor and one of the group factors. In the interest of economy of space, only the 15 phrases which show the highest loading on each factor are reported. A complete list of the loadings of all phrases on all factors appears in the investigator's doctoral dissertation (Bartlett, 1959).

I. *Halo.* This factor shows substantial loadings on almost all of the phrases. Thus, it seems to be a general tendency to rate a person high or low on the basis of some general impression toward the person being rated. The phrases showing the highest loadings on this factor also correlated highly with a rating on overall leadership, indicating that this halo may have been a function of characteristics which are possessed by leaders. For this reason an interpretation of general leadership ability might also be appropriate for this factor. However, since the ratings on over-all leadership were also subject to the halo effect, an interpretation of halo or general bias seems more appropriate.

The fifteen phrases which have the highest loadings on this factor are given with their respective loadings as follows:

Ideas show good judgment (.84)
Ideas are excellent (.83)
Makes many worthwhile comments (.83)
Answers wisely (.82)
Has good judgment (.82)
Gives well thought out answers (.82)
Has many intelligent ideas (.81)
Suggestions are worthwhile (.81)
Has good ideas (.81)
Is a good thinker (.81)
Does important things well (.81)
Give better arguments than others (.80)
Serves as guide to rest of group (.80)
Says things that are constructive (.80)
Is on the ball (.80)

II. *Contribution of Ideas and Information.* This factor seems to be made up of phrases which describe a person who is not only perceived as being intelligent, but also as having a fund of information to contribute to the group. In a classroom discussion situation, the presence of this kind of person seems appropriate.

The 15 phrases with highest loadings on this factor follow:

He knows his stuff (.36)
Very intelligent person (.36)
Knows what he is talking about (.36)
Can grasp essentials of situation (.34)
Gets best test grades in class (.34)
Expresses ideas well (.33)
Thinking is original (.32)
Suggestions are worthwhile (.32)
Well-informed (.31)
Knows his stuff (.31)
Is a real brain (.30)
Able to evaluate suggestions of others (.30)
Has many intelligent ideas (.30)
Has good ideas (.30)
Knows more than others in the group (.29)

III. *Contribution of Friendly Atmosphere.* The phrases loading on this factor seem to describe the type of person who is easy to get along with and who might serve as a morale booster in a group discussion. Although this type of person may not contribute to the goal of the group directly, the friendly atmosphere created may work indirectly in aiding a group to achieve its goals.

The phrases representing this factor with their respective factor loadings are as follows:

Well-liked by all (.57)
Easy to get along with (.55)
Friendly (.55)
Always very pleasant (.54)
Manner is pleasant (.54)
Well-liked by others (.52)
Very friendly (.52)
Seems to have a likeable personality (.51)
Has an agreeable disposition (.50)
Always willing to listen to the other side of an argument (.47)
Always has a smile (.47)
Very agreeable (.47)
Does not overexercise power as a leader (.46)
Does not acquire resentment of others in group (.46)
Always cooperative (.46)

IV. *Contribution of Labor and Effort.* The phrases which make up this factor describe a person who is the "work horse" of the discussion group. This type of person contributes to the group by taking responsibility for such things as writing reports, bringing in material, and taking care of details.

The phrases which represent this factor and their loadings follow:

A hard worker (.65)
More than willing to do his share (.59)
Willing to do extra things (.53)
Takes more than his share of work home (.52)
Always ready to serve (.51)
Willing to work (.50)
Shows a persistent effort (.49)
Not afraid to do more than his share (.49)
Is a good worker (.48)
Seems to enjoy work (.46)
Does not sit by and let others do the work (.45)
Takes responsibility very seriously (.44)
Earnest about work (.43)
Shows enthusiasm in all projects (.41)
Does work outside of class (.41)

V. *Contribution of Policy and Decisions.* The phrases which define this factor describe a person who sees himself as the central figure of the group. This type person pushes the group toward a decision, even at the risk of making it a personal rather than group decision. This factor seems to represent the behavior of a person who might be placed in the traditional category of autocratic leader. The phrases defining this factor follow with their respective loadings:

Often inclined to think he is the only one right (.56)
Insists on his way (.56)
Prefers to do entire project by himself (.49)
Often seems to dominate group discussions (.44)
Doesn't listen to others' opinions (.40)
Not always democratic (.39)
Quite dominant (.35)
Forever involved in a discussion (.35)
Seldom changes his opinion (.35)
Leaves more "timid souls" of group behind (.33)
Argues his point until convinced he is wrong (.33)
Has a solution for every problem (.31)
Argues his point until he convinces others (.30)

Always willing to add an extra word (.27)
Comments arouse controversy (.25)

Discussion

The general factor was expected to emerge from this analysis, since all of the phrases which were used were descriptions of the behavior of persons who were leaders in a discussion group situation. Also in check list rating scales, a general factor would be expected, since response bias is not controlled. Since most of the phrases which have highest loadings on the general factor also have high loadings on the factor indicating a Contribution of Ideas and Information, it would seem that this is the kind of contribution a group member should make if he aspires to being classified as the group leader. The halo in ratings of students in the classroom appears to be set by academic type performance.

The four group factors were interpreted in terms of four different ways of contributing to group activity. These four dimensions of group contribution suggest four independent ways by which a person can become a group leader. The person who could contribute to a group in all of these ways would, presumably, be the more desirable type leader, but such persons are in limited supply. It might be hypothesized that a group made up of four persons, each strong in one of the areas defined by the group factor, would be an efficient group. A rating scale has been constructed using the phrases from this analysis (Bartlett: 1958, 1959) which might be useful to the instructor in setting up the membership in groups in order to test such an hypothesis. However, the use of this rating scale in courses other than educational psychology would be questionable.

In viewing the dimensions found in this study as they are related to dimensions of leadership found by others, two of the group factors from the present study seemed to be similar to factors found in studying leadership behavior in other situations. The two strongest factors (in terms of percentage of total variance explained) of the Ohio State University Leadership Studies, as reported by Halpin and Winer (1952), were Consideration and Initiating Structure. The Consideration factor seems to be similar to the factor indicating Contribution of Friendly Atmosphere, which was found in the present study. The factor, Initiating Structure, appears to be similar to the factor of the present study, Contribution of Policy and Decisions.

The two factors, Contribution of Ideas and Information and Contribution of Labor and Effort, appear to be a function of the classroom situation. In this situation the group orientation is toward learning the course-related materials. Thus, ideas and information play an important role. In these groups of students, the work done in achieving the goals of the group is done by the students who make up the group. Thus, the leader of the group would also be expected to do his share of the work.

Summary

A factor analysis of 300 phrases describing leadership behavior in classroom discussion group was done using the Wherry-Winer Method (1953). The study was done to examine the dimensions of leadership behavior in a classroom group discussion situation. The analysis yielded four group factors and a large general factor. The general factor was interpreted as a general tendency to make high or low applicability ratings of the phrases on the basis of the halo effect. The four group factors were interpreted in terms of the ways which a group member can contribute to the group discussion: ideas and information, friendly atmosphere, labor and effort, policy and decisions.

REFERENCES

Bartlett, C. J. The relationships between self-ratings and peer ratings on a leadership behavior scale. Unpublished doctoral dissertation, Ohio State Univer., 1958.

BARTLETT, C. J. The relationships between self-ratings and peer ratings on a leadership behavior scale. *Personnel Psychol.*, 1959, **12**, 237–246.

HALPIN, A. W., & WINER, B. J. *The leadership behavior of the airplane commander.* Columbus: Ohio State Univer. Res. Found., 1952.

STEAD, W. H., & SHARTLE, C. L. *Occupational counseling techniques.* New York: American Book, 1940.

STOGDILL, R. M. Personal factors associated with leadership: A survey of the literature. *J. Psychol.*, 1948, **25**, 35–71.

WHERRY, R. J. Hierarchial factor solutions without rotation. *Psychometrika*, 1959, **24**, 45–52.

WHERRY, R. J., & WINER, B. J. A method for factoring large numbers of items. *Psychometrika*, 1953, **18**, 161–179.

(Received May 21, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

INDIVIDUAL DIFFERENCES IN MEMORY

JAMES B. STROUD AND LOWELL SCHOER
State University of Iowa

One of the persistent problems in the field of memory has been that of individual differences in retentiveness. Historically, differences in retentive ability have been linked with differences in learning ability. For example, McGeoch and Irion (1952) wrote: "By and large, individual differences in learning are reflected in individual differences in retention."

When subjects have been given equal amounts of practice and amount learned has been used to define differences in acquisition, learning scores have been found to be positively related to retention scores. This is probably what McGeoch and Irion had in mind. Another facet of the problem concerns the relationship between learning and retentive abilities when all *Ss* attain a common performance criterion and differences in the rate at which this is done are used to define differences in learning ability. The evidence pertaining to this aspect of the general problem is less conclusive. It is with this aspect that the present paper is concerned.

The main body of evidence in support of the proposition that retentive ability is positively related to learning ability has been derived from analysis of learning and relearning scores. There is little doubt that there exists a positive and substantial correlation between trials to learn and trials to relearn upon the part of an array of *Ss*. This type of analysis, however, seems to be a questionable procedure. Relearning scores reflect learning ability to some extent, perhaps to a considerable extent. We would expect that *Ss* who learn rapidly would relearn rapidly because, if for no other reason, they are rapid learners. Similarly, those who learn slowly would be expected to relearn slowly, because they are slow learners. At best, relearning scores reflect both learning ability and retentive ability.

On the other hand, the relationship between recall and rate of learning is far from close, when all *Ss* achieve a common trials-to-learn criterion, as Underwood's (1954) data, for example, and those herein reported show. In the customary verbal learning experiment, it is possible for the slow learner to achieve a level of mastery equal to that of the rapid learner if he takes enough time. In many types of learning situations, especially those regarded as most significant educationally, it is, of course, not possible for all learners to attain equal mastery. Some will achieve levels of insight, or levels of skill, not attainable to others. Under these circumstances, we would expect to obtain wide differences among good and poor learners in performance on retention tests.

In the problem at hand, we were interested in individual differences in retention among *Ss*, all of whom have attained a common criterion. Demonstrating that recall scores are unrelated to rate of learning, when all *Ss* achieve the same level of learning, is not, of course, equivalent to demonstrating that significant differences in recall do not exist. A basic question facing us today is whether or not individuals differ reliably among themselves in retention of materials which have been mastered equally well by all. If it could be shown that *Ss* do differ reliably in recall ability and that these differences are not associated with differences in learning ability, then recall, and possibly retention, would be established as an independent variable—especially so if the differences possess some degree of generality.

Procedure

With these thoughts in mind the senior author planned and carried out the experiments reported in the following pages.[1] Originally it was planned to have each of a number of *Ss* learn, recall, and relearn four or five different kinds of material. Later, since it seemed desirable to have the *Ss* learn pairs of comparable lists so that some kind of assessment of the reliability of the scores could be made, this plan was given up. The materials adopted consisted of two lists of paired adjectives, 12 pairs per list, and two lists of picture-name pairs, 10 pairs per list. All lists were presented by the paired-associate anticipatory method.

Adjective pairs of low associative, synonymity, and familiarity value were selected from the Haagen (1943) list. These were arranged in two comparable lists as mentioned. In the preparation of the picture-name pairs, a suitable number of photographs of senior, male students in the College of Engineering were selected from a pool of pictures just taken for the University Student Annual. Paired with each such picture chosen was a fictitious first and last name. The adjective pairs and the picture-name pairs were then reproduced on motion-picture film in such a way that the first member of a pair could be exposed on a screen for two seconds to be followed immediately by a two-second simultaneous exposure of both members of a pair. In the case of the adjective pairs, the second member, the response member, appeared just to the right of the first or stimulus member. In the case of the picture-name pairs the name, the response item, appeared directly under the picture, the stimulus item. An interval of .5 second intervened between exposures of adjacent pairs within a list. An interval of five seconds occurred between trials.

The items in each list were presented in two different serial orders, the two orders alternating in a given series of trials. After the first trial, the *Ss'* task was to anticipate (orally) the response member within the two-second interval in which the stimulus member alone was exposed. This was repeated for each pair within the list, trial after trial, until the learning criterion (a trial without error) was reached.

The items comprising each list, in two different serial orders, were filmed separately, making each of the four reels consisting of some 50 feet of film. Each such length of film was made into a loop and mounted in a winder. The films thus ran continuously during any one learning session for as many trials as required to reach the criterion. Once the running of the film was started for a given *S* the experimenter had no duties beyond recording the responses and stopping the projector at the end of the learning session. The four lists were presented to the *Ss* in a counterbalanced order.[2]

The *Ss* were given five practice trials, on a 6-pair list of adjectives, just prior to the learning of the first paired adjective list. They were given five practice trials on a 4-pair list of picture-names just prior to learning the first picture-name list.

The *Ss* ($N = 149$) were sophomore students enrolled in the Introduction to Psychology course, State University of Iowa. They reported one at a time for five consecutive days, starting on a Monday. On Mondays the *Ss* learned the experimental list appropriate to each for that day. On Tuesdays the *Ss* first recalled and relearned Monday's lists and proceeded to

[1] The author wishes to acknowledge the assistance of the following persons in conducting the experiment: L. J. Carter, A. J. Edwards, Clifford Howe, and S. Muehl.

[2] The method of presenting the materials just described was adopted because of its convenience in handing the picture material. It proved to be quite as satisfactory as a memory drum in handling the verbal materials. The method permits a great deal of flexibility.

learn Tuesday's list. This procedure continued until Friday, on which day the *Ss* only recalled and relearned Thursday's lists.

Results

In Table 1 are presented means of the learning and relearning scores in trials, mean recall scores in words anticipated, and *SDs* of the three respective scores. As may be inferred from the size of the *SDs* of learning scores, the distributions of trials to learn were rather markedly skewed to the right.

As a first step in the analyses of the data, within-orders intercorrelations were computed (*a*) among the learning scores, (*b*) among the relearning scores, and (*c*) among the recall scores. Since we wished to be able to determine practice effects, we presented the lists in counterbalanced order. In order to obviate any possible attentuating effects of this procedure upon the correlations, within-orders correlations were computed, except where otherwise indicated. The results are reported in Table 2. In a sense the correlations involving comparable lists, those in the top two rows, may be regarded as reliability measures. All the *r*'s are significant at the .01 level of confidence. Overall, the *r*'s between various relearning tasks are about .05 higher than those between the various recall tasks. Incidentally, we computed the *r*'s between saving scores for comparative purposes. These were found to be .25 between PA_1 and PA_2, and .11 between PN_1 and PN_2.

Shown in Table 3 are the *r*'s between (*a*) learning and recall, (*b*) relearning and recall, (*c*) learning and relearning, and (*d*) learning and relearning by levels of recall. The latter in effect is the correlation between learning and relearning with recall partialled out. The *r*'s between learning and recall are not significant in the case of the two paired adjectives lists, but are significant at the .01 level on the picture-names lists. At best the relation-

TABLE 1

Mean Trials to Learn and Relearn, Mean Recall Scores, and *SDs* of Paired Adjectives and Picture-Names Lists

	PA_1	PA_2	PN_1	PN_2	*N*
Learning[a]	11.85	10.50	13.03	14.33	149
SD	7.57	5.86	6.05	7.36	149
Recall	8.20	8.40	6.09	5.55	149
SD	2.19	2.10	2.18	2.32	149
Relearning	3.54	3.17	3.52	3.68	149
SD	1.75	1.84	1.70	1.73	149

[a] Does not include first presentation trial.

TABLE 2

Within-Orders Correlations among Variables

Learning	Relearning	Recall
PA_1, PA_2 .69	PA_1, PA_2 .34	PA_1, PA_2 .23
PN_1, PN_2 .79	PN_1, PN_2 .34	PN_1, PN_2 .30
PA_1, PN_1 .46	PA_1, PN_1 .29	PA_1, PN_1 .41
PA_1, PN_2 .44	PA_1, PN_2 .38	PA_1, PN_2 .27
PA_2, PN_1 .46	PA_2, PN_1 .32	PA_2, PN_1 .34
PA_2, PN_2 .45	PA_2, PN_2 .40	PA_2, PN_2 .24

Note.—r .159 (N = 150) significant at .05 level; r .206 (N = 150) significant at .01 level. Values are averages, based upon z transformations, of the r's of the four orders.

TABLE 3

Within-Orders Correlations among Learning, Relearning, and Recall Scores

Scores	Tasks			
	PA_1	PA_2	PN_1	PN_2
Learning and Recall	.04	−.12	−.25	−.23[a]
Recall and Relearning	−.40	−.59	−.34	−.30[b]
Learning and Relearning	.34	.40	.43	.55
Learning and Relearning by levels of recall (average)[c]	.39	.37	.37	.48

[a] Learning and relearning scores are trials scores. Thus negative *r*'s between these scores and recall signify positive relationships.

[b] Not within-orders.

[c] Values are based upon weighted averages of z transformations of the r's for the separate levels.

ship between learning and recall, as obtained, is quite low. This may be to some extent a function of relatively low reliability of recall scores.

As an additional attempt to assess the relationship between learning and recall scores, the *Ss* were divided upon the basis of level of learning scores into fifths on each task and mean recall scores determined for the respective fifths. These were found to be as follows:

List	F_1*	F_2	F_3	F_4	F_5
PA_1	8.7	8.2	8.3	8.4	7.6
PA_2	8.2	8.8	8.6	8.2	8.9
PN_1	5.3	6.2	6.7	6.0	6.4
PN_2	5.0	5.9	6.1	5.1	5.8

* The slowest fifth.

None of the differences among these means for any given list departs significantly from zero. In this treatment, recall and learning scores certainly appear to be essentially unrelated.

As seen in Table 3, significant correlations were obtained between recall and relearning scores for each of the four lists of material. This is a logical outcome. Even if recall and learning ability are not in fact related at all, we should expect to obtain significant correlationships between recall and relearning because the *Ss* who recall most have least to relearn. Indeed this suggests that relearning, at least after a 24-hour interval, is a valid measure of retention. Our earlier discussion has not denied this. We maintained only that it is not known to what extent relearning scores represent retention and to what extent, learning ability.

It seemed logical to us that by computing correlations between learning and relearning by levels of recall, that is, by partialling out recall effects, we could obtain some idea of the extent to which relearning scores are affected by recall. We did this. The results are reported in Table 3. Overall, the partialling out procedure resulted in slightly lower correlations. Perhaps this is what one should expect in the light of the *r*'s obtained between learning and recall. However, the distribution of the *r*'s for the various levels of recall is quite revealing. The magnitude of the *r*'s varied inversely with the level of recall, as is illustrated by the values for lists PA_1 and PA_2 combined, as follows, starting with the lowest level:

Level of Recall	Total *N*	*r* Learning & Relearning
1	30	.65
2	28	.64
3	45	.46
4	62	.39
5	46	.34
6	34	.08
7	15	.00

Our data, together with those of workers in the past, suggest at best no more than a slight positive relationship between rate of learning (to a common trials criterion) and recall. Underwood (1954) has made the point that the attaining of a common trials criterion upon the part of fast and slow *Ss* does not guarantee equal habit strength by the two groups. At least it probably does not do so for all of the items in a list. This, if true, does not forbid the use of designations fast and slow upon the basis of a trials criterion nor the making of comparisons of retention of the two groups thus designated. However, Underwood's point is highly significant for the basic question of individual differences in retention among *Ss* who do in fact achieve equal mastery in acquisition by a habit strength criterion.

By a special treatment of his data, Underwood compared fast learners (upper 50%) with slow learners (lower 50%) on one day recall of items on which the two groups, during acquisition trials, had shown equal probabilities of correct responses on subsequent trials, once correct responses had been made. The two groups did not differ in recall of the items thus defined. We analyzed our results for five levels of learning performance. Of course, the *Ss* did not during the acquisition trials manifest equal probability of making correct responses on subsequent trials once

correct responses had been made. This inequality is probably one of the basic differences between fast and slow learners, as others have suggested. Perhaps *Ss* differ in powers of discrimination and differentiation which permit them to make correct responses in the first place. This may be regarded as one basis for differences in rate of acquisition. Another basis, as suggested by inference, is differences in the degree to which reinforcement (or at least the making of a correct response) contributes to habit strength. Cutting across, or interacting with, both are differences in susceptibility to intraserial interference, as illustrated in length-of-list effects. For example, Carter (1958) has shown that the increasing of length of lists affects slow learners more adversely than fast learners. Indeed, differences in susceptibility to intraserial interference, as between fast and slow learners, may account for the observed differences in effectiveness of reinforcement.

We computed (*a*) the number of reinforcements taken to learn each list by each fifth of the *Ss*, by rank order, (*b*) the number of item presentations to reach the criterion, and (*c*) the ratio of number of item presentations (P) to the number of reinforcements (R). The results are presented in Table 4.

TABLE 4

NUMBER OF REINFORCEMENTS (R) BY *Ss* PERFORMING AT VARIOUS ACQUISITION RATES, NUMBER OF ITEM PRESENTATIONS (P), AND RATIO OF P TO R

	Mean Trials to Learn by Fifths	Mean Number Item Presentations (P)	Mean Number Reinforcement (R)	R/P
		PA_1		
5	5.86	70.32	40.37	.57
4	8.77	105.24	60.43	.57
3	10.80	129.60	77.73	.60
2	14.07	168.84	99.67	.59
1	23.73	284.76	162.67	.57
		PA_2		
5	5.57	66.80	38.27	.57
4	7.93	95.16	57.50	.60
3	10.40	124.80	76.97	.62
2	13.40	160.80	93.30	.58
1	18.66	223.92	140.25	.63
		PN_1		
5	7.47	74.70	38.47	.51
4	10.83	108.30	55.33	.51
3	12.93	129.30	68.27	.53
2	16.13	161.30	81.23	.50
1	22.90	229.00	118.47	.52
		PN_2		
5	7.48	74.80	34.89	.47
4	10.55	105.50	51.66	.49
3	14.14	141.40	68.79	.49
2	17.83	178.30	91.13	.51
1	25.43	254.30	126.97	.50

It is seen that the number of reinforcements involved bears a more or less constant ratio to the number of item presentations at all five levels of learning performance, the mean value of R/P for the four lists combined being .53, .54, .55, .54, and .56. Perhaps this ratio would vary for all ability levels, as it seems to do so here, with the over-all difficulty of the list. This tells us that the mean number of reinforcements, in an array of *Ss*, bears a fairly constant ratio to the number of item presentations or, to what amounts to the same thing, the number of trials. The mean number of reinforcements accomplished by the slowest fifth of the *Ss* is 3.6 times as great as that by the fastest fifth. Another way of stating the matter is that the *Ss* in the top fifth gained on the average .29 of an item per reinforcement, while those in the lowest fifth gained only .08 of an item per reinforcement, averages for all four lists, or 44 items. The middle fifth of *Ss* gained .15 of an item per reinforcement, by the same method of computation. A similar value for the second highest fifth is .20 and for the second lowest fifth, .12.

In the course of attaining the criterion of one perfect trial, the slowest fifth of

the *Ss* had 3.6 times as many reinforcements as the fastest fifth. If we allow that the fastest learners have some small over-all advantage in recall over the slowest learners, it would follow that the 3.6 to 1 ratio in reinforcements is not quite sufficient to insure equal habit strength. For our two extreme groups and for the two adjacent groups, here referred to as Fast, Moderately Fast, Moderately Slow, and Slow, we computed the probability of correct responses occurring on the next trial after varying numbers of reinforcements. These results for the two lists of paired adjectives combined are shown in Table 5.

Taking these results at their face value, we may observe that the slowest group of *Ss* required approximately five reinforcements to establish a probability of a correct response occurring equal to that established by the fastest group after one reinforcement. The probabilities for the moderately slow group after three reinforcements and for the moderately fast after two, approach the probabilities of the fast group after one reinforcement. For all four groups two consecutive reinforcements seem to result in about the same probabilities as three reinforcements without regard to consecutiveness.

TABLE 5

PROBABILITIES OF CORRECT RESPONSES FOLLOWING VARYING NUMBERS OF REINFORCEMENTS FOR *Ss* REPRESENTING FOUR LEVELS OF ACQUISITION RATE

Reinforcement	Slow		Mod. Slow		Mod. Fast		Fast	
	T[a]	Prob.	T	Prob.	T	Prob.	T	Prob.
1	475	.549	329	.647	489	.755	489	.890
2	472	.674	316	.788	465	.871	385	.945
3	455	.796	294	.881	416	.913	247	.980
4	440	.854	270	.926	344	.939	138	.986
5	414	.894	294	.940	265	.988	42	.976
6	394	.898	221	.955	189	.995	22	1.000
Consecutive Reinforcement								
2	467	.806	288	.872	439	.932	362	.964
3	405	.894	259	.950	370	.938	320	.983
4	376	.915	249	.960	296	.942	125	.984

[a] T—total number of possibilities. i.e., for the slow group there were 475 instances in which these *Ss* had a chance to make a correct response following one reinforcement; 472 instances in which there was a chance following two reinforcements, etc.

In the normal course of learning, there will be some number of instances in which the slowest *Ss* cannot for given times establish a 5 to 1, 4 to 1, etc. ratio of reinforcements in a relationship to the fastest group of *Ss* because of the termination of the experiment. It is possible, though not highly probable, that there will be some instances in which, for the slowest learners, items will be reinforced but once. There should be a greater number in which but two or but three reinforcements occur. Logically, these should be poorly retained, relatively. Indeed this condition might be sufficient to account for any small difference in recall between fast and slow learners. Among our slowest *Ss* we found 24 instances in which there had occurred but 1, 2, or 3 reinforcements. The 24-hour recall scores of such items by the *Ss* involved was 37.5%, as compared with 70% of all the items in the two lists by all *Ss* in the slowest group. The items involved in this treatment were pretty well distributed over the entire lists. Obviously little can be concluded from such meager data.

It is suggested that when *Ss* learning at different rates develop equal habit strength they should not differ, by acquisition rates, in retention, unless the inappropriate method of relearning is used as the measure of retention. However, the usual method of requiring all *Ss* to learn to a trials criterion does not quite guarantee equal habit strength. In such cases the rapid learners may show a slight advantage in retention. Conceivably, this might vary with the rigorousness of the criterion to which the *Ss* were required to learn.

Incidentally, if rapid learners behave on long lists as slow learners do on short lists, as Carter's work suggests, then we might infer that *Ss* at any given acquisition level should retain lists of varying lengths equally well when learned to a common habit strength criterion.

This leaves open one of the major questions raised above: do individuals differ among themselves in retention of material originally learned to the same level by a habit strength criterion? Our *Ss* all learned to a criterion of one perfect trial, but not to the same habit strength criterion. The six intercorrelations among the respective recall scores varied from .23 to .41, all significant at the .01 level. However, in the case of the PN_1 and PN_2 lists, significant correlations were obtained between learning and recall.

As an attempt to partial out the effects attributable to differences in rate of acquisition, *r*'s were computed between recall scores on these two tasks by levels of learning ability. In doing this, the learning scores on the two tasks, PN_1 and PN_2, were averaged and eight levels of learning rate established. The average *r* by *z* transformation was .15. The results of the various analyses performed leave this issue in doubt.

Finally, for the data at hand, we sought to determine practice effects on learning and on recall. The means of the scores by orders in which the performances occurred are presented in Table 6. As seen, practice affected rate of acquisition, as was to be expected, but did not affect recall to any appreciable degree.

TABLE 6
PRACTICE EFFECTS UPON LEARNING AND RECALL

Day	Mean Trials to Learn[a]				Mean Recall			
	PA_1	PA_2	PN_1	PN_2	PA_1	PA_2	PN_1	PN_2
1	13.28	12.31	14.79	17.17	8.93	9.27	6.29	6.66
2	10.69	11.15	12.51	13.05	8.38	8.15	5.78	5.14
3	10.38	9.39	12.18	13.81	8.14	9.46	6.83	6.12
4	10.56	9.40	11.19	12.38	7.56	7.33	5.31	4.65

[a] Does not include first presentation trial.

SUMMARY

This paper has recounted some of the salient features of an investigation of individual differences in memory. One hundred forty-nine *Ss*, in individual sessions, learned and recalled and relearned four lists each, two of paired adjectives and two of paired picture-names. Significant correlations, ranging from .23 to .41, were obtained among the various recall scores. On the two paired adjectives lists, *r*'s between trials to learn and words recalled (after 24 hours) were not significant. Significant *r*'s of −.23 and −.25 were obtained between these variables on the two picture-names lists. *Ss* were divided into fifths upon the basis of trials to learn on each of the four lists. No significant differences were obtained among the recall scores of the various fifths on any of the lists. At best, the results of the various analyses suggest no more than a slight relationship between rate of learning and recall.

The data have been analyzed with respect to differences in habit strength among *Ss* of varying rates of learning, all of whom attained a common trials-to-learn criterion. It was found that the mean number of reinforcements bore a fairly constant ratio to the number of item presentations (or number of trials) for the different levels of learning ability. The mean number of reinforcements accomplished by the slowest fifth of *Ss* was 3.6 times as great as that by the fastest fifth. Further analysis of specific items showed that the slowest fifth required five reinforcements to establish a probability of making a correct response on the succeeding trial equal to that established by the fastest fifth after one reinforcement. The important question of whether or not *Ss* who have learned to a common habit strength criterion differ significantly

among themselves in retention is unanswered. The data are in accord with the proposition that significant differences in retention do exist among *Ss* who have achieved a common trials-to-learn criterion.

Practice effects reduced the trials required to learn, but did not result in increased recall scores.

REFERENCES

CARTER, L. J. Interrelationships among memory, rate of acquisition, and length of task. Unpublished doctoral dissertation, State Univer. of Iowa, 1958.

HAAGEN, C. H. Learning and retention as a function of the synonymity of original and interpolated tasks. Unpublished doctoral dissertation, State Univer. of Iowa, 1943.

MCGEOCH, J. A., & IRION, A. L. *The psychology of learning.* New York: Longmans, Green, 1952. P. 325.

UNDERWOOD, B. J. Speed of learning and amount retained: A consideration of methodology. *Psychol. Bull.*, 1954, **51**, 276–282.

(Received June 8, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

A NOTE ON THE EFFECT OF FILLING OUT AN "ANXIETY SCALE" ON EXAMINATION PERFORMANCE

NEIL A. CARRIER

Southern Illinois University

Certain statistically significant relationships were found in an earlier study by the writer (Carrier, 1957) designed to determine whether certain personality measures are related to course examination performance under stress conditions. But the conclusions appeared tenuous because mean score differences between high and low stress experimental groups were generally small. *Post hoc* speculation suggested several tenable explanations. One of these involved the old problem of the extent to which the investigator's measuring operations themselves may unwantedly influence that which is measured.

As a part of the study procedures, a brief "scale" had been administered to assess the subjects' (*Ss*') awareness of anxiety or tenseness both before and during the examination. It is possible that in attempting to measure the *Ss*' awareness of anxiety, the measuring instrument itself had increased the anxiety responses of the *Ss* through some kind of "feedback" or suggestion process. Further, if some differential effect were present whereby the low stress *Ss* increased more in anxiety than did the already highly anxious high stress *Ss*, the result would have been to reduce the desired difference in stressfulness between the high and low stress groups. As a consequence (providing stress does affect examination performance), there would have been less difference in examination performance between the two groups than would otherwise have been found.

A study by Coppock (1955) added further credance to this possibility. Drawing support from other sources (Horney, 1937; Mowrer, 1950) he proposed that an awareness of one's emotional reactions might thereby increase the amplitude or duration of such reactions. His study's results gave some support to the hypothesis that "suggestion and stress can influence the extent of an individual's overreaction to information about his own autonomic responses" (Coppock, 1955, p. 28). Since the present author's earlier study attempted to get self-reports of autonomic responses, one effect of the measure designed to get these reports may have been to increase emotional responses where it was especially not wanted, i.e., in the low stress group.

Consequently, a partial "replication" was planned to determine whether filling out anxiety scales does in fact depress examination performance.

PROCEDURE

The *Ss* were 1074 students in five lecture sections of the introductory psychology course at the University of Colorado during the fall 1956 semester. Two forms of a 75 item, multiple-choice examination were administered. Both forms contained the same 75 questions, but the order of questions was different in the two forms. The two forms were alternated in distribution to the students.

One form was arbitrarily designated as the "anxiety" form, the other was called the "neutral" form. Two pages, each containing seven graphic rating scales, were included with the anxiety form. The first page appeared at the beginning of the exam. It requested the *S* to "place a large X at some place along each line which will

best indicate the strength of your feeling on each question." The first question asked, "How Tense or Anxious about this exam do you feel at this moment?" At one end of a six-inch line was the phrase "very tense," at the other end appeared "very relaxed"; "midpoint" appeared in the middle. The remaining questions and scales asked how Well Prepared the student felt for the exam, to what extent he was aware of an "Uneasy Feeling," how Confident he felt he would do well on the exam, the extent to which he noticed an Accelerated Heartbeat at the moment, whether he was Perspiring more than usual at the moment, and to what extent he felt Worried about the exam.

Approximately in the middle of the exam the second page of seven graphic rating scales appeared. Printed instructions were to "indicate your feelings at this point in the examination. Do not refer to your answers to the scales at the beginning of the exam." The questions and scales here were the same as those presented at the beginning of the examination.

The second, or neutral, form also consisted of two pages. The first page appeared at the beginning of the exam. It requested the following information: name, sex, age, year in school, local address, and home state. The second page appeared approximately in the middle of the exam. It asked for "the following information before continuing this exam: name, city of birth, state of birth, mother living, father living, number of brothers, and number of sisters." The neutral form sheets obviously were intended to occupy an amount of time and attention of the control group of *Ss* comparable to that demanded from the experimental group of *Ss*, without, however, confronting the former with anxiety-producing materials.

Results and Discussions

The mean examination score for the 553 students who filled out the anxiety scales was 53.33; the mean score for the 521 students who filled out the personal data sheets was 53.36. The standard deviation of the former group was 8.89; that for the latter group was 8.95. Both distributions were normal in shape by visual inspection.

These remarkably similar results give no support to the contention that filling out anxiety scales during an examination depresses performance by "reminding" the student of his emotional state. Whether there was an increase in emotional reactions as a result of the self-rating procedures was, of course, not directly assessed. But no effect on examination performance was detected. It may be added that had a direct influence been found, research of this kind would be more difficult than it already is.

It appears, then, that one interpretation of the small score differences between high and low stress groups found in the earlier study has been discounted. Remaining alternative explanations (among others) point to a failure to get a large "spread" on stress between the two experimental groups (probably, to lower the stress in the low stress group), or to examination stress having less effect upon performance than is popularly claimed.

REFERENCES

Carrier, N. A. The relationship of certain personality measures to examination performance under stress. *J. educ. Psychol.*, 1957, **48**, 510–520.

Coppock, H. W. Responses of subjects to their own galvanic skin responses. *J. abnorm. soc. Psychol.*, 1955, **50**, 25–28.

Horney, Karen. *The neurotic personality of our time*. New York: Norton, 1937.

Mowrer, O. H. *Learning theory and personality dynamics*. New York: Ronald, 1950.

(Received June 24, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

PSYCHOLOGICAL HEALTH AND CLASSROOM FUNCTIONING:

A STUDY OF DISSATISFACTION WITH SCHOOL AMONG ADOLESCENTS[1]

PHILIP W. JACKSON AND JACOB W. GETZELS
University of Chicago

The problem of dissatisfaction with school among children is of theoretical and practical significance to both psychologists and educators. At the theoretical level dissatisfaction with school becomes part of a broader area of inquiry which aims at an understanding of the individual's functioning in an institutional setting and which includes studies of staff morale, role conflict, productivity, and the like. At a practical level the question of why children like or dislike school is directly related to the immediate problems of school dropouts, grouping procedures, planning for the gifted child, and the like.

As might be expected, a social phenomenon as important as dissatisfaction with school is not without its explanatory hypothesis. Some of these spring from empirical findings, while others appear to be part of our cultural ethos. Educational studies that point to an empirical linkage between school failure and school dropouts, and industrial studies that demonstrate a relationship between low morale and decreased output, lead one to suspect that reduced effectiveness in school (i.e., low scholastic achievement) would be a natural concomitant of dissatisfaction with the institution. Thus one would expect to find heightened dissatisfaction among students who have low ability or who are unable for one reason or another to deal adequately with scholastic material.

More recently it has been suggested (although never adequately demonstrated) that many successful students with high ability are dissatisfied with their school experiences; the term "boredom" is often linked with the term "gifted child" in current expositions by educators. The boredom problem among "gifted" combined with the failure experiences of the low ability child suggests that the greatest number of dissatisfied students is to be found among extreme ability groups. Those who are low in ability and achievement would be expected to show dissatisfaction because of the numerous frustrations they experience in the classroom. Those who are high in ability and achievement would be expected to show dissatisfaction because of the relative lack of stimulation which they experience in the classroom.

Both of these explanations (or, more accurately, hypotheses) contain the implication that dissatisfaction with an institution arises out of the individual's interaction with that institution. An alternative explanation might be that the individual brings a set toward satisfaction or dissatisfaction *to* the institution—that it is a reflection of a more pervasive personal orientation and that success or failure experiences within the institution have a limited influence upon it. This hypothesis obviously places more emphasis than do the earlier ones upon psychological variables, as opposed to environmental variables, in understanding dissatisfaction with school. The research described here was designed to test the relative merit of these alternative views.

[1] This study was supported by a research grant from the United States Office of Education. The present report is an expanded version of a paper read at the American Psychological Association meeting, Cincinnati, Ohio, September 1959.

Problem

The purpose of this investigation is to examine the differences in psychological functioning and classroom effectiveness between two groups of adolescents—those who are satisfied with their recent school experiences and those who are dissatisfied.

Subjects and Procedure

The *Ss* of this investigation were two groups of adolescents identified from among 531 students enrolled in a Midwestern private school. These students were divided into five class groups ranging from the prefreshmen to the senior year of high school. In this institution a single grade, the prefreshmen, is substituted for the usual seventh and eighth grades. The instrument used to select the experimental groups, called the Student Opinion Poll, was a 60-item opinionnaire designed to elicit responses concerning general satisfaction or dissatisfaction with various aspects of school—viz., the teachers, the curriculum, the student body, and classroom procedures. The following are sample items, one in each of the four areas.

3. While there are some differences among them, most teachers in this school are:
   a. Very inspiring
   b. Quite inspiring
   c. Somewhat inspiring
   d. Not inspiring

16. Most of the subjects taught in the school are:
   a. Interesting and challenging
   b. Somewhat above average in interest
   c. Somewhat below average in interest
   d. Dull and routine

14. From the standpoint of intellectual ability, students in this school are:
   a. Too bright—it is difficult to keep up with them
   b. Just bright enough
   c. Not bright enough—they do not provide enough intellectual stimulation

5. The freedom to contribute something in class without being called upon by the teacher is:
   a. Discouraged more than it should be —students do not have enough opportunity to have their say
   b. Encouraged more than it should be —students seem to be rewarded just for speaking even when they have little to say
   c. Handled about right

The instrument was scored by giving one point each time the *S* chose the "most satisfied" response to a multiple-choice item. Thus, the possible range of scores was from 0 to 60. For the total school population the mean score on the Student Opinion Poll was 37.30; the standard deviation was 9.57. The experimental groups were chosen as follows:

Group I—the "dissatisfied" group—consisted of all students whose score on the opinionnaire was at least one and a half standard deviations *below* the mean of the entire student body. This group contained 27 boys and 20 girls.

Group II—the "satisfied" group—consisted of all students whose score on the opinionnaire was at least one and a half standard deviations *above* the mean of the entire student body. This group contained 25 boys and 20 girls.

The experimental groups were compared on the following variables:

1. *Individual intelligence tests.* In most cases this was the Binet. A small number of children were given the Henmon-Nelson, the scores of which were converted by regression equation into equivalent Binet scores.

2. *Standardized verbal achievement test.* The Cooperative Reading Test was used. Prefreshmen and freshmen were given Test C_1, Form Y; older students were given C_2, Form T.

3. *Standardized numerical achievement tests.* Because of differences in the curricula of the various grade groups it was not possible to administer the same test of numerical achievement to all *Ss*. The following tests were given according to grade placement:
   Prefreshman—Iowa Everypupil Arithmetic Test, Advanced Form O.
   Freshmen—Snader General Mathematics Test.
   Sophomores—Cooperative Elementary Algebra Test, Form T.
   Juniors—Cooperative Intermediate Algebra Test.
   Seniors—Cooperative Geometry Test, Form 2.

4. *California Personality Test.* Two forms of this instrument were used. The intermedi-

ate form was given to prefreshmen; the secondary form was given to all of the older groups. Two subscores were obtained, "personal adjustment" and "social adjustment."

5. *Direct Sentence Completion Test.* *Ss* were asked to complete 27 sentences of the type: "When I saw I was going to fail I .....," or, "I think my father is ....." Each sentence was given a plus or minus score depending upon the presence or absence of morbid fantasy, defeatism, overt aggression, and the like. The total score was the summation of the individual sentence scores.

6. *Indirect Sentence Completion Test.* This instrument was identical with the Direct Sentence Completion Test except that proper names were inserted for the pronoun "I," thus changing it from a "self-report" to a "projective" instrument. Boys' names were used in the male form of the instrument and girls' names in the female form. The instrument was presented as a "thinking speed" test. To reinforce this notion *Ss* were asked to raise their hands when they were finished and the elapsed time was written on their test booklet. This instrument was administered approximately two weeks prior to the administration of the Direct Sentence Completion Test.

7. *Group Rorschach.* Cards III, IV, IX, and X were projected on a screen. For each picture the *S* was presented with 10 responses and was asked to choose the three which he thought to be most appropriate. Each list of 10 contained four "pathological" responses. The *S*'s score was the number of nonpathologic responses among his 12 choices. This group technique follows that described by Harrower-Erikson and Steiner (1945).

8. *Teacher ratings.* Each student was given three ratings by his present teachers. These ratings included: (*a*) his general desirability as a student; (*b*) his ability to become involved in learning activities; and (*c*) his possession of leadership qualities. Teachers were required to place all of their students on a five-point scale so that Categories 1 and 5 each contained one-twelfth of the students; Categories 2 and 4 each contained one-fourth of the students; and Category 3 contained one-third of the students. The values 5, 8, 10, 12, and 15 were assigned to the categories and were used in quantifying the ratings.

9. *Adjective Check List.* From a list of 24 adjectives each student was asked to choose the 6 which best described his characteristic feelings while attending classes in particular school subjects. The list contained 12 "positive" (e.g., confident, happy, eager, relaxed) and 12 "negative" adjectives (e.g., bored, restless, misunderstood, angry). The use of the negative adjectives by the experimental groups was analyzed both quantitatively and qualitatively.

Results

With the exception of the adjective check list the results of all comparisons are shown in Table 1. Contrary to popular expectations the "satisfied" and "dissatisfied" students did *not* differ from each other in either general intellectual ability or in scholastic achievement. Those differences which did appear were linked to psychological rather than scholastic variables. More specifically, each of the test instruments designed to assess psychological health or "adjustment" was effective in distinguishing "satisfied" from "dissatisfied" students within one or both sex groups.

For both sexes the experimental groups were differentiated by their scores on the California Test of Personality. The experimental groups of boys were further differentiated by their responses to the Indirect Sentence Completion Test. For girls additional differences appeared in their responses to the Direct Sentence Completion Test and the Group Rorschach.

On all of these test variables the "satisfied" group attained the "better" score—i.e., the score signifying a more adequate level of psychological functioning. It is also worthy of note that whenever a significant difference appeared, the mean score of the total student population fell between the mean scores of the experimental groups. Thus, the variables that differentiate the experimental groups tend also to distinguish them from the total population of students.

In addition to showing differences on psychological health variables, "satisfied" and "dissatisfied" boys were perceived differently by their teachers. On all three of the teachers' ratings the "satisfied" boys

TABLE 1

MEAN SCORES, STANDARD DEVIATIONS, AND *t* STATISTICS FOR SATISFIED AND DISSATISFIED ADOLESCENTS ON DEPENDENT VARIABLES[a]

	Boys					Girls				
	Dissatisfied (N = 27)		Satisfied (N = 25)		t	Dissatisfied (N = 20)		Satisfied (N = 20)		t
	$\bar{x}$	s	$\bar{x}$	s		$\bar{x}$	s	$\bar{x}$	s	
IQ	134.85	14.58	136.44	14.59	ns	128.45	15.06	128.00	11.45	ns
Verbal Achievement	49.96	8.69	50.68	7.87	ns	50.63	9.11	52.28	6.76	ns
Numerical Achievement	50.35	9.75	52.17	10.52	ns	47.78	8.61	48.50	10.26	ns
Calif. Personal Adjust.	45.58	9.82	53.40	7.63	3.18**	47.90	13.03	54.76	9.25	1.86*
Calif. Social Adjust.	44.85	11.37	51.84	8.93	2.45**	47.00	13.15	55.76	7.89	2.50**
Direct Sentence Comp.	46.93	10.58	49.25	10.02	ns	46.65	12.01	54.00	5.73	2.53**
Indirect Sentence Comp.	47.19	9.61	51.29	6.95	1.75*	49.60	10.35	53.47	7.97	ns
Group Rorschach	48.35	10.66	47.44	10.30	ns	47.35	11.35	54.16	8.32	2.15**
Teacher Rating I: Desirability as a student	8.94	1.83	10.35	1.70	2.85**	9.84	1.91	10.05	1.59	ns
Teacher Rating II: Leadership qualities	9.01	2.08	10.13	1.96	2.00*	9.91	2.37	10.04	1.24	ns
Teacher Rating III: Involvement in learning	9.09	2.14	10.23	1.69	2.14**	9.67	2.32	10.33	2.11	ns

* Significant at the .05 level.

** Significant at the .01 level.

[a] With the exception of IQ, all scores were based upon parameters of the total student body from which the experimental groups were drawn. The scores of all tests were transformed to T scores with a mean of 50 and a standard deviation of 10. For the total population the teacher ratings have a mean of 10 and a standard deviation of 2. The mean IQs for the total school population are: boys, 132, and girls, 128.

TABLE 2

NUMBER OF SUBJECTS CHOOSING NEGATIVE ADJECTIVES WHEN ASKED TO DESCRIBE TYPICAL CLASSROOM FEELINGS

Adjective	Boys			Girls		
	Dissatisfied (N = 27)	Satisfied (N = 25)	Chi Square	Dissatisfied (N = 20)	Satisfied (N = 20)	Chi Square
Inadequate	19	16	ns	17	7	10.42**
Ignorant	19	13	ns	15	3	14.54**
Dull	25	16	6.36*	16	9	5.60*
Bored	24	13	8.61**	20	13	8.48**
Restless	20	15	ns	19	9	11.90**
Uncertain	20	21	ns	17	13	ns
Angry	15	4	8.76**	13	4	8.29**
Unnoticed	19	5	13.25**	7	4	ns
Unhelped	18	8	6.24*	9	6	ns
Misunderstood	16	5	8.31**	5	2	ns
Rejected	12	3	6.66**	4	0	ns
Restrained	17	2	16.91**	9	3	4.29*

* Significant at the .05 level.

** Significant at the .01 level.

received more favorable judgments than did "dissatisfied" boys. The fact that this result does not appear to be true for girls lends support to the popular expectation that boys are more likely to express their negative feelings publicly than are girls. This hypothesis receives some confirmation from the results of the adjective check list which are described below.

In Table 2 are shown the number of *Ss*

who chose negative adjectives when asked to describe their typical classroom feelings. As they are arranged in Table 2 the adjectives reflect the rankings of four judges who were asked to rank the words on the degree to which they involved an implicit or explicit criticism of others. The 12 adjectives were typed on separate cards and were accompanied by the following directions:

> On the following cards are a number of negative adjectives which a person might use to describe himself. Rank these adjectives on the degree to which they involve an implicit or explicit criticism of others. For each adjective ask the question: If a person used this adjective *to describe himself* would he also be implicitly or explicitly criticizing others? Give a rank of 1 to the adjective which would be *least* critical of others and a rank of 12 to the adjective which would be *most* critical of others.

Four psychologists served as judges. The average rank order correlation among the four sets of judgments was .84. The adjectives are presented in Table 2 according to the ranked sum-of-ranks of the judges. The adjective "inadequate" was judged as being most free of criticism of others, while the adjective "restrained" was judged as involving the greatest amount of criticism of others.

As might be expected, the use of negative adjectives was far more frequent among dissatisfied students than among satisfied students. Four adjectives seemed to discriminate equally well between the experimental groups for both sexes; these were: "bored," "angry," "restrained," and "dull."

An examination of Table 2 also suggests the existence of sex differences in the students' description of their typical classroom feelings. Remembering the classificatory scheme by which the adjectives are ranked in Table 2, it appears that dissatisfied girls are somewhat less likely than dissatisfied boys to use negative adjectives involving implicit criticism of others. Dissatisfied boys, on the other hand, are less likely than dissatisfied girls to be distinguished from their satisfied counterparts by the use of adjectives *not* involving implicit criticism of others. If one thinks of criticism directed towards others within Rosenzweig's schema of "intropunitiveness" and "extrapunitiveness" (Murray, 1945), then the observed sex differences may be conceptualized by saying that dissatisfied girls are more *intropunitive* than satisfied girls; dissatisfied boys are more *extrapunitive* than satisfied boys.

This difference in the direction of aggression may provide a context for the obtained differences in teacher ratings discussed earlier. If the dissatisfied boy is more likely than his female counterpart to lay the blame for his dissatisfaction upon others in his environment, particularly school authorities, it is reasonable to expect that he would be viewed as somewhat less than completely desirable by the classroom teacher. The dissatisfied girl, on the other hand, seems more willing to direct her negative feelings inward, thus avoiding the additional risk of counter-aggression by school authorities or by other adults.

Discussion

Two major conclusions are suggested by the findings of this study. First, dissatisfaction with school appears to be part of a larger picture of psychological discontent rather than a direct reflection of inefficient functioning in the classroom. It is almost as if dissatisfaction were a product of a pervasive perceptual set that colors the student's view of himself and his world. Second, it appears that the "dynamics" of dissatisfaction operate differently for boys and girls. Boys seem to project the causes of their discontent upon the world around them so that adults are seen as rejecting and lacking in understanding. This tendency to blame adults may be one reason why these boys are seen as less attractive by teachers than are satisfied boys. Girls, on the other hand, are more likely to be

self-critical, turning blame for their dissatisfaction inward. Feelings of inadequacy, ignorance, and restlessness more sharply differentiate satisfied and dissatisfied girls than is the case with boys. This tendency to be intropunitive may partially explain why teacher ratings fail to distinguish between our two experimental groups of girls.

The atypicality of the sample population used in this research places a number of limitations upon the inferential statements which can be made on the basis of these findings. Fortunately, however, the major portion of the investigation has recently been replicated using seventh and eighth grade lower-class Negro adolescents as *Ss* (Spillman, 1959). The findings of the latter study are essentially the same as those reported here. Again the psychological rather than the intellectual or scholastic variables discriminated between satisfied and dissatisfied students. The findings with respect to the use of negative adjectives were not as clear-cut but, again, every intropunitive adjective was used more frequently by dissatisfied girls as compared with dissatisfied boys, while the latter exceeded the girls in their use of extrapunitive adjectives.

It should be noted that even the most satisfied students made some use of negative adjectives when asked to describe their typical feelings in the classroom. Also, the average member of the satisfied group expressed some dissatisfaction on one-sixth of the questions in the Student Opinion Poll. These two observations should serve as ample cautions against the danger of interpreting any sign of dissatisfaction with school as symptomatic of deeper psychological difficulties. Apparently, some degree of dissatisfaction is the rule rather than the exception. Nonetheless, the responses of the extremely disgruntled group of students leaves little doubt that dissatisfaction with school, like beauty, is frequently in the eye of the beholder.

Summary

This investigation examines the differences in psychological functioning and classroom effectiveness between two groups of adolescents—those who are satisfied with their recent school experiences and those who are dissatisfied. The major findings point to: (*a*) the relevance of psychological health data rather than scholastic achievement data in understanding dissatisfaction with school; (*b*) the importance of differentiating the attitudes of dissatisfied girls from those of dissatisfied boys, the former being characterized by feelings of personal inadequacy, the latter by feelings critical of school authorities. Rosenzweig's concepts of intropunitiveness and extrapunitiveness are applied to these findings and a relevant theoretical framework is proposed.

REFERENCES

Harrower-Erikson, M. R., & Steiner, M. E. Large scale Rorschach techniques. Springfield, Ill.: Charles C Thomas, 1945.

Murray, H. A. *Explorations in personality*. New York: Oxford Univer. Press, 1938.

Spillman, R. J. Psychological and scholastic correlates of dissatisfaction with school among adolescents. Unpublished master's thesis. Univer. of Chicago, 1959.

(Received July 6, 1959)

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 6, 1959

NOTE TAKING DURING OR AFTER THE LECTURE

SIGMUND EISNER AND KERMIT ROHDE
Oregon State College

Though note taking during lecture is strongly advocated by most how-to-study books, one is hard put to find a theory which would advocate the note-taking process as a means of learning in and of itself. The resulting notes themselves, of course, probably are in most cases better for study, and hence retention, than no notes at all. But the question remains whether the student cannot use his time during the lecture for some more beneficial activity. If the student takes notes *after* the lecture, he is freed during lecture for such activity *and* induced to perform such activity in order that he might make notes, —and still he has notes available for later study. Furthermore such a procedure forces an immediate attempt to recall. Spitzer (1939) found, working with prose passages, that an immediate attempt to recall improved retention, and Gates (1917), as a result of his retention studies, suggested that taking notes after lecture might be more conducive to retention than taking them during lecture.

Problem

The problem was to test the hypothesis that taking notes during a lecture is a less effective method of retaining lecture material than listening to the lecture and taking notes immediately afterward.

Subjects

A class of 60 students in a college English Literature course at the freshman level was used as subjects.

PROCEDURE

On the basis of the instructor's grades and impression during the previous three weeks, members of the class were paired for ability. The class was then split into two groups, the only stipulation being that each group contain one member of each pair.

Two 30-min. lectures were given to the class: one a lecture on the short story, and one about four weeks later on romanticism. During the first lecture, Group I was told to take notes during the lecture and then to study them for the 15 min. immediately following, and Group II was told to take no notes during the lecture, but to listen and be prepared to take notes and study them immediately after the lecture. To prevent the complicating effects which might arise from varied amounts of additional study, the notes were collected after the 15-min. period following the lecture. This same procedure was followed during the second lecture except that Group I took notes after and Group II during the lecture.

The specific instructions were given as follows: (The material in brackets was deleted at the second lecture period and that in parentheses was deleted at the first lecture period.)

> Today we are going to conduct (the second part of) an experiment on note taking (in order to provide a control). (You will remember) What we wish to find out is whether it is more effective to take notes during lecture or immediately after lecture. In a few minutes, therefore, Dr. E. will divide the class into two groups, one of which is to take notes during the lecture as usual, and the other is to *concentrate intensely on the lecture* while it is being given and then immediately after the lecture jot down all the notes you can. (Each of you will be in the opposite group from last time.) After Dr. E. has divided the class, he will give the lecture—a 30-min. lecture on the [short story] (novel)—and then there will be a 15-min. period of silence in order for you to jot down and study your notes, or just to study your notes, depending upon which group you are in. Then you will sign and

hand in your notes, which we will study for the purposes of our experiment and return to you at a later date. Part of the next meeting will be devoted to a test covering this material. Though the test results will not affect your grade, please try to do your best on this test. Are there any questions before we divide the class?

The names in each group were then read and the lecture given. After the lecture, Dr. E. said:

All right, those of you who have notes study them, and the others jot down any notes you wish, and study them.

At the end of the 15 min. Dr. E. said:

Will you please sign your notes and hand them in?

Two days later, the class was given a 50-item true–false test on the lecture and an essay question with the following instructions:

This is the test I mentioned at our last meeting. Though it will not count on your grade, please do your best.

For the lecture on romanticism the essay question was:

You have heard a lecture which defines romanticism. According to this lecture what are the romantic elements in *Huckleberry Finn*? (The class had just finished reading *Huckleberry Finn*.)

For the short story lecture, the question was:

If fiction writers are interested in truth, and if fiction is the opposite of fact, in that fact has occurred while fiction was invented, why do seekers of truth write fiction?

Three weeks after the second lecture, the class was given an unannounced test (hereinafter referred to as the "late" test) containing 100 true–false items, 50 on the short story lecture and 50 on the romanticism lecture.

All true–false tests were scored in terms of numbers of errors; and, with the names removed so one could not tell from which group they came, the essays on each question were ranked by the lecturer according to the degree they showed a grasp of his material.

Present on all of the five days during which this experiment was in progress were 24 members from Group I and 22 members from Group II. One member of Group I was dropped because he skipped a page of the test. Another member chosen at random was eliminated from Group I in order to make it equal in size to Group II.

Results

Three tests were made with the data.

First for each student his score on the true–false items covering the lecture during which he took notes was compared with his score on the items covering the lecture after which he took notes. The results are given in Table 1.

Since no significant differences appeared between simultaneous and delayed note taking on either the early or late test, it was decided to check on the possibility that a given note taking procedure would favor a student at a high level of ability but impede a student of a low level. The students were divided into two such groups on the basis of their end of term grade in the course. This procedure placed into each of the experimental groups 10 good (A and B) students and 12 poor students. Again no significant differences between simultaneous and delayed note taking were found in either the early or the late test within either the good or the poor student group.

Secondly, to check on the possibility that certain kinds of items were better retained under one method than the other, a difference score between the immediate note takers and the delayed note takers was computed for each item. This was done using the 44 *Ss* appearing for all parts of the experiment except for the first short story test. Since there were 27 *Ss* in each group who had heard this lec-

TABLE 1

ERRORS ON TRUE-FALSE TEST

Test	All Students ($N = 44$)			Good Students ($N = 20$)			Poor Students ($N = 24$)		
	M	SD	t[a]	M	SD	t	M	SD	t
Early:									
Notes during lecture	11.4	4.4	0.90	9.2	4.8	1.43	13.3	3.1	0.24
Notes after lecture	12.0	3.7		10.9	3.7		13.0	3.5	
Late:									
Notes during lecture	14.0	4.6	0.00	11.8	3.9	0.73	15.9	4.2	0.72
Notes after lecture	14.0	4.1		12.5	3.2		15.3	4.3	

[a] Individuals were paired with themselves for the t test.

ture and taken the early test, all 54 were used for this analysis.

Of the 200 items, only 10 showed significant differences at the .05 level between immediate and delayed note takers as determined by the Chi square test. Since one would expect by chance exactly 10 questions to show such differences, this line of investigation was concluded. It was noted, however, that the five questions on which the immediate note takers were superior were the more general, and the five on which the delayed note takers were superior were the more detailed questions; a trend contrary to what one might expect.

Finally, the 54 essay answers on the early short story test, with the names removed, were ranked by the lecturer according to the degree of understanding of the lecture evidenced in the essay. Next the 44 essay answers on the early romanticism test were similarly treated. Then this data was analyzed by the Mann-Whitney test to determine whether the delayed note takers remembered the kind of information necessary to write such a test better than the immediate note takers. In both tests the immediate note takers ranked higher. But as the value for U in the first test was only 383 (mean $U = 365$, $\sigma U = 57.8$) and in the second test only 263 (mean $U = 242$, $\sigma U = 42.6$), neither value approached significance at the .05 level of confidence.

DISCUSSION

When one considers that note taking has, with respect to most students, an advantage of many years of practice over any other method, it is surprising to find that note taking during lecture in and of itself fails to show any superiority to delayed note taking. Past experience was not without effect as a number of students in the present experiment were very upset at not being allowed to take notes during lecture even though grades were not involved. The question suggests itself if they had been similarly experienced in delayed note taking, would this method have shown itself superior as most theory would suggest.

The speculation that, at least for the student who has a great deal of difficulty concentrating on a lecture, note taking is an aid is not supported in this study. One would expect students at this level of interest to be the poorer students, but the results for the poorer students taken separately did not favor immediate note taking.

SUMMARY

Students in a college English literature class were given two lectures. During one they took notes and studied afterward; during the other they listened and then took notes and studied afterward. Several days later they were given the "early"

test of true–false items and an essay question; three weeks later the "late" test of true–false items only. The data on 44 students gave no significant differences between the two methods on either the early or the late true–false test (either for the total group or for good and poor students taken separately). Nor did the comparison of the essay tests from the "during" and "after" groups yield a significant difference. Hence, no evidence was found to support the belief that note taking itself during lecture is any more effective than note taking immediately after lecture.

REFERENCES

GATES, A. I. Recitation as a factor in memorizing. *Arch. Psychol.*, 1917, **6**, No. 40.
SPITZER, H. F. Studies in retention. *J. educ. Psychol.*, 1939, **30**, 641–656.

(Received May 14, 1959)

GRADUATE EDUCATION IN PSYCHOLOGY

Report of the Conference on Graduate Education in Psychology, sponsored by the Education and Training Board of the American Psychological Association and supported by a grant from the National Institute of Mental Health, U. S. Public Health Service; held at Miami Beach, Florida, November 29 to December 7, 1958.

Prepared by the Editorial Committee:

Anne Roe, *Chairman*,

and

John W. Gustad, Bruce V. Moore,

Sherman Ross, and Marie Skodak

Price $1.50

THE AMERICAN
PSYCHOLOGICAL ASSOCIATION
Dept. Grad
1333 Sixteenth Street, N.W.
Washington 6, D. C.

Again Available

SCHOOL PSYCHOLOGISTS AT MID-CENTURY

Edited by

NORMA E. CUTTS

The Report of the Thayer Conference, published in 1955, covered an intensive study of the development of school psychology. Because of continuing demand, the Report has been reprinted.

230 pages. Price $2.75

AMERICAN PSYCHOLOGICAL ASSOCIATION
1333 Sixteenth Street, N.W.
Washington 6, D. C.